데이터 웨어하우스 환경에서 데이터 품질의 향상을 위한 개념적 프레임워크의 개발에 관한 연구

정 경 수* · 김 병 곤** · 장 상 도***

A Study on the Development of Framework for Enhancing Data Quality in Data Warehouse Environments

Kyung Soo Chung · Byung Gon Kim · Sang Do Jang

요 약

데이터 웨어하우스와 데이터 품질에 관한 문헌연구를 통하여 데이터 웨어하우스 환경에서 데이터 품질의 향상을 위한 개념적 프레임워크를 개발하고자 하는 것이 본 연구의 목적이다. 데이터 웨어하우스 데이터 품질향상 활동을 지원하는 프레임워크를 개발하는 목적은 (1) 다양한 요구를 가진 사용자들이 웨어하우스 데이터에 접근하기 때문에, 사용자의 요구를 만족시키며 기업의 목적에 적합한 품질향상 활동을 지원하기 위해서이며, 다양한 기업활동을 가장 잘 지원할 수 있는 데이터 품질향상 지침을 관리자에게 제공하기 위해서이다. (2) 웨어하우스 관리자의 데이터 품질향상 활동을 지원하기 위해서는 품질차원이나 데이터세트 등과 같은 품질향상에 필요한 다양한 이슈를 관리자가 인식할 수 있도록 하기 위해서이다. (3) 데이터 웨어하우스 환경에서 데이터 품질 향상에 필요한 체계적이고 포괄적인 안목을 제공하기 위해서이다.

본 연구는 다음과 같은 단계로 수행하게 된다. 첫째, 데이터 웨어하우스의 개념과 데이터 웨어하우스의 구축단계 및 데이터 웨어하우스를 구성하는 프레임워크를 검토한다. 둘째, 데이터 웨어하우스 환경에서의 데이터 품질의 기준과 데이터 품질의 측정 및 데이터 품질의 향상 방안 등을 고찰한다. 셋째, 데이터 웨어하우스 환경에서 데이터 품질의 향상을 위한 개념적 프레임워크를 개발하기 위하여 데이터 웨어하우스 데이터 품질 향상과 관련된 기업활동, 데이터 세트, 품질의 속성 및 차원 등을 정의한다. 마지막으로 데이터 웨어하우스 환경하에서 데이터 품질을 향상할 수 있는 3차원 구조의 개념적 프레임워크를 제안하며, 나아가 제안한 모형에 대하여 데이터 품질 향상을 위한 프로젝트 활동의 사례를 통하여 모형의 타당성을 개념적으로 설명한다.

1. 서 론

오늘날 우리의 기업들은 치열한 경쟁에서 살아남기 위하여 많은 노력을 기울이고 있다. 그러나 지난 몇 년 사이에 기업을 둘러싼 급격한 환경의 변화를 살펴보면 가히 혁명적이라 할 수밖에 없으며 이러한 환경의 변화에 능동적으로 대처하기 위해서는 신속한 의사결정이 필요하며 미래의 시장환경을 예측하기 위한 정확하고도 관련성이 높은 정보가 요구된다.

지금까지 기업에서는 정보시스템을 개발하고 데이터베이스를 구축하여 업무의 신속한 처리와 함께 의사결정에 필요한 정보를 확보하는데 많은 노력을 기울여 왔다. 그러나 기존의 데이터베이스는 기업에 필요한 자료를 저장하고 관리하는 차원에서 주로 현재의 상황을 나타내주는 자료들이 대부분이었으며, 또한 기업내 여러 부문에서 독립적인 관리가 이루어지는 것이 일반적이었다. 그러나 이러한 환경하에서는 시계열 데이터의 저장이나 엄청난 분량의 데이터를 효과적으로 분석하거나 관리하는 것이 쉽지 않으며, 효율적인 정보관리는 이루어지기 어렵다.

최근 이러한 문제점들을 극복하고 효과적인 조회와 분석을 통하여 의사결정의 질을 높이고 기업의 경쟁력을 강화하려는 데이터 웨어하우징의 개념이

* 경북대학교 경영학부 교수
** 안동과학대학 사무자동화과 교수
*** 경북대학교 대학원 경영학과

소개되었다. 기존의 데이터베이스를 의사결정 차원에서 한 층 더 끌어올린 것이 데이터 웨어하우징이며, 데이터베이스에 저장되어 있는 모든 데이터들이 사용자의 의사결정에 영향을 주지는 않으므로, 정확하고, 완전하며, 시기적절한 데이터들을 활용함으로써 기업의 의사결정에 도움을 주려는 것이 데이터 웨어하우징의 목적이다. 현재까지 데이터 마이닝을 비롯한 여러 가지의 정보분석 도구들이 소개되고 있는데, 이러한 도구들이 데이터 웨어하우스에 결합되어 기업내의 의사결정에 활용이 되며 정보관리의 중요한 역할을 담당하게 된다.

그러나 데이터 웨어하우스에는 여러 가지의 원천으로부터 장기간에 걸친 엄청난 분량의 데이터와 요약정보가 저장되므로 자료분석도구 뿐만 아니라 저장된 자료의 품질이 중요한 의미를 가지게 되며 의사결정의 결과도 데이터 웨어하우스에 저장된 자료의 품질에 의해 좌우된다. 데이터 산업이 활성화됨에 따라 불량데이터(poor data quality)로 인한 문제도 점점 증가하고 있다. 컴퓨터에 저장된 미국의 범죄관련 자료는 50-80%가 부정확하거나, 불완전하거나, 모호한 것으로 알려지고 있으며[Strong, Lee, and Wang, 1997], 몇몇 사례연구에서 조사한 바에 의하면 기업에서 저장하고 있는 자료의 0.5% 내지 30%가 불량한 자료로 나타났다[Redman, 1998].

이러한 불량자료는 고객불만족, 운영비용의 증가, 비효과적인 의사결정의 증가를 가져오며 기업의 전략수행에 큰 장애를 초래하게 된다. 또한 불량데이터는 종업원의 사기를 저하시키고 조직내 불신을 조장시키는 결과를 낳게 된다. 따라서 데이터의 품질을 향상시킨다는 것은 그 자체로도 경쟁에서 살아남기 위한 기업의 목표 중의 하나가 될 수 있으며[Diane, Yang, and Richard, 1997; Thomas, 1998] 특히 데이터웨어하우징 환경에서 데이터의 품질은 데이터웨어하우스의 성과와 가치에 결정적인 영향을 미친다 [Ballou and Tayi, 1999].

본 연구에서는 데이터 웨어하우스 환경하에서 데이터의 품질을 향상시키기 위한 프레임워크를 개발하고 개념적인 논의를 하였다. 본 논문의 구성은 서론에 이어 2장에서는 데이터 웨어하우스의 개요에 대하여 논의하였으며, 3장에서는 데이터의 품질에 대하여 선행연구들을 중심으로 논의하였다. 이어 4장에서는 데이터 품질향상을 위한 개념적 프레임워크를 소개하였으며, 마지막 5장에서는 결론을 제시하였다.

2. 데이터 웨어하우스의 개요

2.1 데이터 웨어하우스의 개념

데이터 웨어하우스에 대해서 많은 학자들이 다양한 관점에서 데이터 웨어하우스를 정의하고 있다. 우선 Inmon[1996]은 "기업 의사결정 과정을 지원하기 위한 주제 중심적이고, 통합적이며, 시간성을 가지는 비휘발성 자료의 집합"이라고 정의하였다. 이것은 운영시스템 및 운영용 데이터베이스와의 차이를 중심으로 정의한 것으로 가장 일반적으로 인용되고 있다. Arun과 Varghose[1998]는 의사결정지원시스템의 기초로 이용되는 읽기 전용의 데이터베이스라고 정의하였으며, Ballou와 Tayi[1999]는 기업내에서 다양한 플랫폼 및 아키텍처 상에서 구현된 다양한 데이터 모델을 포함하고 있는 문제를 해결하기 위한 주제지향적 전사적 데이터베이스라고 정의하고 있다. 지금까지 검토한 문헌을 중심으로 데이터 웨어하우스의 특징을 살펴보면 다음과 같다.

1) 주제지향적 : 데이터가 일정한 주제별로 모아져야 한다는 것이다. 예를 들어, 은행업무에서 기존의 데이터베이스가 대출, 예금, 은행카드, 신탁처리 등과 같은 응용과 운영들을 중심으로 설계되는 반면에, 데이터 웨어하우스에서는 고객, 거래처, 상품, 활동 등과 같은 주요 주제 영역을 중심으로 조직화된다. 둘의 차이는 데이터의 내용에서 뚜렷하게 나타나는데, 데이터 웨어하우스에서는 의사결정과 무관한 데이터를 포함하지 않는 반면 운영시스템은 의사결정 지원이 아닌 거래처리요구를 만족시키기 위한 운영계의 세밀한 정보를 포함한다.

2) 통합성 : 데이터 웨어하우스 내의 모든 데이터는 항상 통합되어 있어야 한다는 것으로 일관된 이름짓기방법, 일관된 변수측정, 일관된 코드화구조, 일관된 데이터의 물리적 특성들을 나타내야 한다.

3) 시계열성 : 데이터 웨어하우스에 있는 모든 데이터는 일정 시간 동안에는 정확하다. 즉 운영중심 환경에서는 한 단위의 데이터에 접근할 때 그 데이터가 정확한 값을 가지고 있다는 사실은 접근하는 순간에만 확신할 수 있으나, 데이터 웨어하우스의 데이터는 순간적인 것이 아니라 일정 기간 동안 그 정확성을 유지할 수 있다는 것이다.

4) 비휘발성 : 운영데이터 시스템 환경에서는 레코드 단위를 기초로 하여 삽입, 삭제, 변경 등의 갱신이 이루어지는데 반하여, 데이터 웨어하우스는 초기의 데이터 적재와 접근의 두 가지 활동만을 수행하므로 데이터의 갱신은 이루어지지 않는다[정희원, 1997; 정명호, 1998]. <표 1>은 운용 데이터베이스와 데이터 웨어하우스의 차이점을 비교한 것이다[조재희, 1996; Bontempo & Zagelow, 1998].

<표 1> 운용데이터베이스와 데이터 웨어하우스의 차이점

구 분	운용 데이터베이스	데이터 웨어하우스
구성의 특성	• 업무처리(transaction)중심 업무별 데이터 집합 • 의사결정에는 필요하지 않더라도 업무처리에 필요한 데이터는 모두 관리대상	• 분석(analysis) 중심 • 주제별 데이터 집합 • 의사결정에 필요한 데이터만 유지
자료의 통합	• 통합되지 않음	• 통합 (속성의 이름, 자료의 표현, 도량형의 단위 등이 일관성이 있음. 표현, 단위등이 데이터 웨어하우스에서는 분석 및 비교를 위해 일관적)
자료의 역사성	• 접근하는 현재 시간(current)을 기준으로 하여 최신의 값을 유지	• 시간에 따라 모든 순간(과거/현재)의 값을 유지
자료의 갱신성	• Volatile(자주 갱신)	• Non-volatile(갱신이 없음)
규 모	• Variable	• Increase only
주요도구	• DBMS, OLTP	• OLAP
취합의 정도	• 상세	• 요약/상세

2.2 데이터 웨어하우스의 유형

1) 집중화된 데이터 웨어하우스(Centralized data warehouse)

대부분의 조직에서는 단일의 집중화된 데이터 웨어하우스 환경을 구축하고 유지한다. 단일의 집중화된 데이터 웨어하우스 환경을 구축하는 것이 바람직한 이유는 다음과 같다[Bontempo & Zagelow, 1998]. 그리고 집중화된 데이터 웨어하우스의 형태를 그림으로 나타내면 <그림 1>과 같다.

(1) 웨이하우스의 데이터는 조직전반에 걸쳐 통합되어지며, 통합된 관점에서 사용되어지는 데이터는 웨어 하우스 본부에만 존재한다.

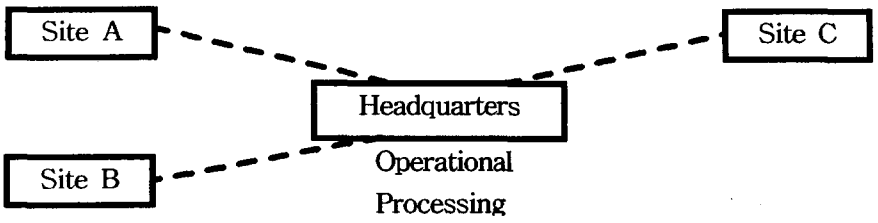
(2) 데이터 웨어하우스에 존재하는 대량의 데이터는 하나의 집중된 데이터 저장소에 저장하는 것이 바람직 하다.

(3) 데이터를 통합하더라도 데이터가 다수의 로컬(local)에 걸쳐 분산되어 있다면, 데이터를 액세스하는데 장애요인이 된다.

(4) 규모의 경제의 이익을 실현할 수 있다.

2) 분산 데이터 웨어하우스(Distributed data warehouse)

분산 데이터 웨어하우스의 경우 많은 처리가 근거리에서 직접적으로 이루어지며, 운영처리가 독립적으로 이루어진다. 그리고 자료처리의 형태에 따라 데이터나 활동들이 센터로 보내진다. 또한 분산 데이터 웨어하우스는 네트워크로 연결된 데이터 웨어하우스를 구성하는 것으로 볼 수도 있다. 즉, 사용자들이 어느 지역에 있든지 간에 집중화된 데이터 웨어하우스에 접근을 하여 필요로 하는 정보를 얻을 수 있다는 것이다. 분산 데이터 웨어하우스의 형태를 나타내면 <그림 2>와 같다.



<그림 1> 집중화된 데이터 웨어하우스

<그림 2> 분산 데이터 웨어하우스

2.3 데이터 웨어하우스의 구축단계

데이터 웨어하우스를 어떻게 구축하느냐 하는 것은 데이터의 품질에도 많은 영향을 준다고 알려져 있는데 여러 가지 구축방법론이 소개되고 있으나, 본 연구에서는 Bort[1999]의 3단계 구축방법론을 중심으로 설명하였다.

1단계 : Pre-engineering

데이터 웨어하우스를 통해 얻고자 하는 목표가 무엇인지 정의한다. 이것은 전 단계를 통해서 가장 중요한 부분으로 확실한 목표가 확립되지 않으면 데이터 웨어하우스는 무용지물이 되고 말 것이다. 또 그러한 목표를 데이터 웨어하우스가 어떻게 지원할 것인지, 데이터는 어떻게 저장이 되고, 필요한 기술적 사항들은 무엇이며, 예산은 어느 정도가 될 것인지 등을 결정해야만 한다.

2단계 : Engineering

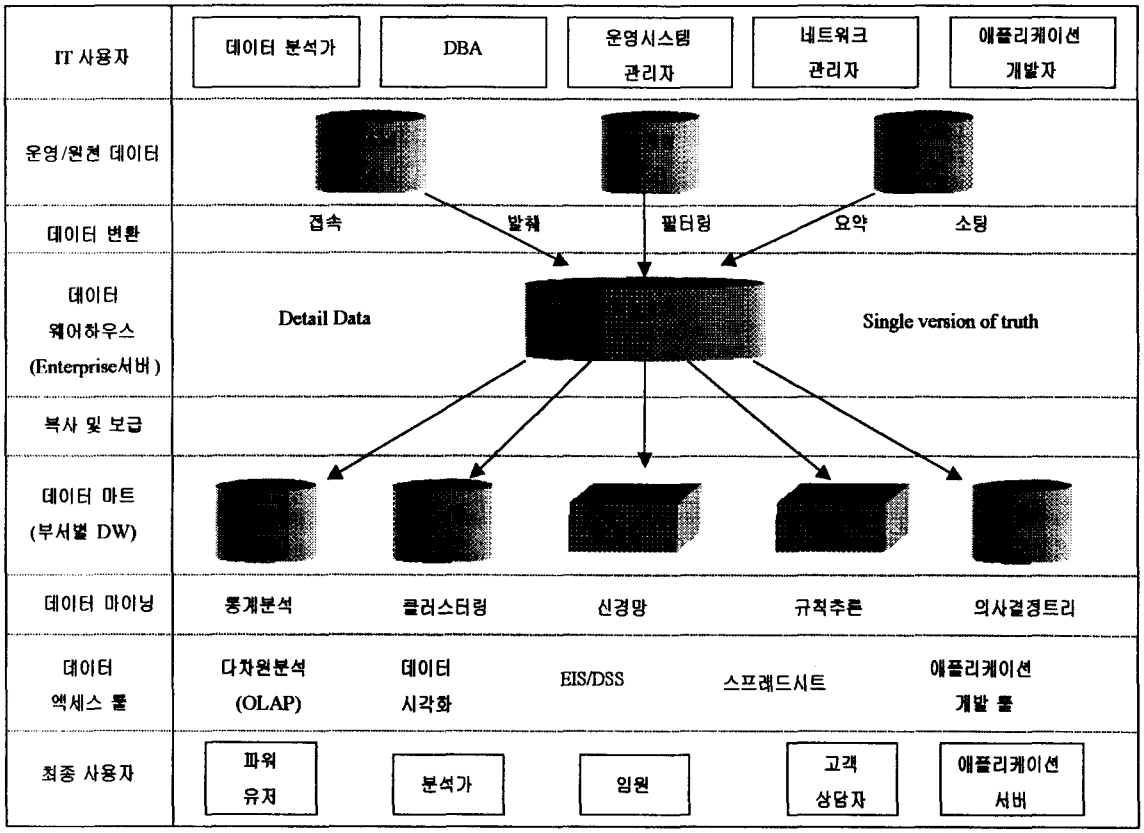
잘 정의된 설계를 가지고 개발자들은 하드웨어와 운영시스템을 선택하고 그에 따른 의사결정을 수행한다. 하드웨어의 사양은 어떻게 될 것인지, 웨어하우스를 구축하느냐 데이터마트를 구축할 것이냐 같은 의사결정이 이에 해당된다. 현재의 추세는 일반적으로 데이터마트를 먼저 개발한 뒤 웨어하우스로 통합하는 방법을 사용하고 있다. 데이터 마트란 송수신 시간을 단축하기 위해 데이터 웨어하우스에서 특별히 관련된 정보만을 별도의 부서나 워크그룹용 서버에 저장하는 것으로 데이터 웨어하우스 정보의 하부세트를 말한다[Bontempo & Zagelow, 1998]. 다음으로 어떤 데이터베이스를 사용할 것인지를 결정한다. Red Brick, Informix, Sybase, Oracle 등과 같은 회사의 제품들이 주로 많이 사용되는 데이터베이스이다.

3단계 : Development

먼저 프로토타입을 통해 원형을 만든다. 그리고 사용자들로 하여금 잘못된 부분과 잘된 부분을 찾게 하는데 처음에는 시각적인 측면을 테스트하고, GUI Interface가 잘 갖추어져 있는지를 살펴본다. 그리고 나서 프로토타입의 결과를 가지고 최소한의 성능을 가진 모델을 만들고, 회사 내에 훈련된 사용자들(power users)에게 제품의 오류를 찾고, 제품의 향상방안과 새로운 아이디어를 제시하게 한다. 마지막으로 피드백을 계속 반복한 뒤 완성 제품을 만든다. 프로토타이핑 방법을 이용하는 것은 개발시간을 단축시키고, 정보시스템의 질을 향상시키며, 경제성이 있다는 장점을 가진다. 또한 사용자 참여를 촉진하여 진정한 요구사항을 짧은 시간안에 결정할 수가 있다[이병수, 이상락, 장근, 1999; Bort, 1997]

2.4 데이터웨어하우스의 개념적 프레임워크

데이터 웨어하우스가 무엇인지를 전체적으로 인식하며, 데이터 웨어하우스에 대한 개념을 보다 명확하게 파악하기 위하여 데이터 웨어하우스의 개념적 프렘임워크를 나타내면 <그림 3>과 같다[조재희, 1997; Gardner, 1998]. 최종사용자는 그들의 업



<그림 3> 데이터 웨어하우스의 개념적 프레임워크

무를 수행하기 위해 원시 데이터를 OLAP 데이터베이스에서 가지고 오지만, 의사결정의 효율성을 위해 데이터 웨어하우스라고 불리는 분석용 DB를 별도로 구축해야 한다. 이 데이터 웨어하우스는 일정한 주기로 계속 과거 데이터를 축적하기 때문에 그 용량이 엄청나게 커야한다는 한계점으로 인해 데이터마트라는 도구를 새로이 구축하고 있으며, 이 방대한 정보들을 분석하는 방법으로서 데이터마이닝 기법이 대두되고 있다. 여기서 데이터 마이닝이란 기업이 보유하고 있는 데이터에서 통계적 방법이나 모델링 기법을 이용하여 기존에 발견되지 않았던 숨겨져 있는 정보를 발견해 내는 과정을 말한다[Ballou & Tayi, 1999].

3. 데이터웨어하우스 환경에서의 데이터품질

데이터베이스와 데이터 웨어하우스의 차이점은 앞에서 설명하였다. 그러면 데이터 웨어하우스가 저장하고 있는 데이터는 데이터베이스가 저장하고 있는 데이터와 차이가 있다는 것을 알 수 있을 것이다. 데이터 웨어하우스는 의사결정과 관련된 형태로 데이터를 저장하고 있다. 즉 원시데이터(source data)는 동일하나 주제중심적으로(의사결정에 유용하도록) 통합되고, 정제되고(refined), 정화된(cleaned) 데이터를 저장하고 있으므로, 양자의 데이터 품질에 대한 접근도 달라야 할 것이다.

서론에서 우리는 의사결정의 질을 높이는 것과 데이터 품질을 향상시키는 것과는 직접적인 관련이 있다라고 설명하였다. 그리고, 데이터의 품질을 향상시키지 못하고 불량한(poor) 데이터를 가졌을 경우 기업에 미치는 영향에 대해 좀 더 구체적으로 설명을 하면 <표 2>와 같다[Thomas, 1998].

본 연구의 목적인 데이터 품질을 향상시키기 위한 방안을 설명하기 위해 먼저 데이터 품질의 개념을 살펴보고, 데이터가 적절한 것인지를 결정하도록 하는 품질의 차원들에 대해 알아 보고자 한다.

데이터 품질이란 데이터 자체의 바람직한 정도를 의미한다[데이터베이스월드, 1996]. 데이터 웨어하우스가 성공을 하지 못하는 이유는 다음의 두 가지로 요약할 수 있다[Ballou & Tayi, 1999]. 첫째로 데이터 품질에 대한 관심의 부족이며, 둘째는 적절

<표 2> 불량한 데이터가 기업에 미치는 영향

운영상의 영향	전술상의 영향	전략상의 영향
- 고객 만족을 저하시킨다. - 비용을 증가시킨다. - 종업원 만족을 저하시킨다.	- 의사결정의 질이 저하된다. - 웨어하우스 구현이 어렵다. - 리엔지니어링이 어렵다. - 조직간 불신이 증가한다.	- 전략 설정이 어려워진다. - 전략 실행이 어려워진다. - 데이터 소유에 관한 문제 야기 - 조직간 협력을 어렵게 한다. - 경영자 관심을 다른 곳으로

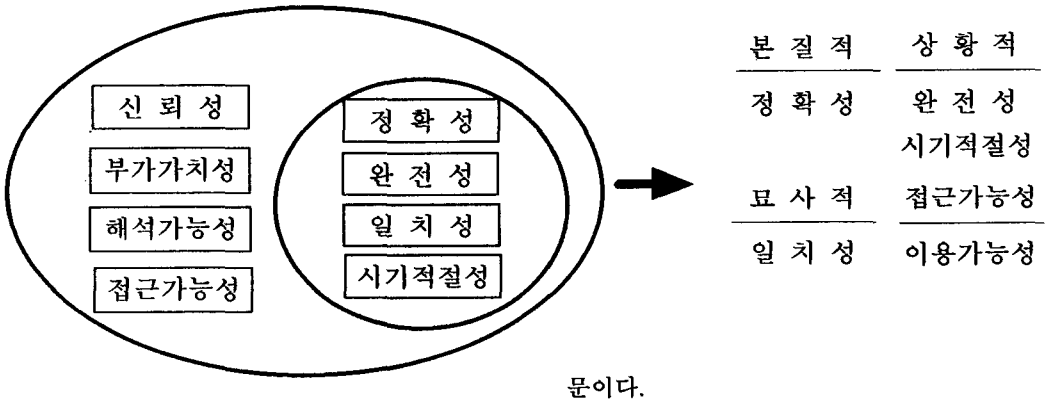
한 데이터를 저장하지 못했기 때문이다. 그러나 두 번째 이유의 경우는 상대적인 개념이 들어가 있다. 즉 어떤 환경에서는 적절할 수도 있지만, 다른 환경에서는 그 데이터가 부적절할 수도 있다는 것이다. 데이터 품질을 "사용의 적합성(fitness for use)"이란 말로 표현하는 이유도 바로 거기에 있다 [Ballou & Pazer, 1998].

Ballou와 Pazer[1985]는 데이터 품질의 기준으로 4가지 차원을 정의했다. 정확성(accuracy), 완전성(completeness), 일치성(consistency), 시기적절성(timeliness), 그리고, 최근에는 Wang과 Strong[1996]이 데이터를 사용하는 사람들의 관점에 따라 신뢰성(believability), 부가가치성(value added), 해석가능성(interpretability), 접근가능성(accessibility)과 그 외의 몇 개를 추가하였다. 그리고 이러한 차원들을 본질적(intrinsic)-정확성, 상황적(contextual)-완전성, 시기적절성, 묘사적(representational)-일치성, 접근가능성(accessibility)-이용가능성 등으로 그룹화 하였다 [Ballou & Tayi, 1999]. 위에서 논의한 Ballou와 Pazer[1985], Wang과 Strong[1996], Ballou와 Tayi[1999] 등이 주장한 데이터 품질의 차원을 종합적으로 나타내면 <그림 4>와 같다.

정확성이란 레코드에 저장하고 있는 데이터와 실제로 존재하는 원시데이터 사이에 의미 또는 내용에 차이가 없는가를 의미하는 것으로 데이터 자체에 오류가 있거나 데이터를 표현하는 과정에서 오류가 발생할 수 있다. 완전성은 가공된 데이터가 원문(원시데이터)에 담긴 정보를 완전하게 담고 있는지를 의미하며, 일치성은 데이터베이스내 둘 이상의 데이터 값이 서로 상충되지 않고 일관된 상태를 이루는 정도로 정의된다. 불필요하게 중복되는 속성이 없어야 하며, 데이터를 요약하고 발췌하는 방식이 또한 일치해야 한다. 모든 레코드의 정보가 일치성을 가지고 있다면 소수의 필드만으로도 원하는 정보를 쉽게 얻을 수 있다. 시기적절성은 사용자가 요구하는 시간에 해당하는 적절한 정보가 신속하게 사용자에게 제공되어야 한다는 것이다.

Yair와 Richard[1996]는 데이터 웨어하우스의 품질의 차원을 '접근가능성', '해석가능성', '시기적절성'의 3가지로 구분하였다. 접근가능성은 의사결정자가 그 데이터를 이용할 수 있어야 한다는 것이며, 해석가능성은 그 데이터를 의사결정자가 쉽게 이해해야 한다는 것을 말한다. 그리고, 시기적절성은 의사결정에 사용되는 데이터는 현재의 상황에 적절한, 시간적 간격을 가지지 않은 것이어야 한다는 것이다.

본 연구에서는 데이터 웨어하우스 환경에서 데이터 품질의 향상을 위한 품질차원으로 '정확성', '완전성', '시기적절성'의 3가지 차원을 이용하고자 한다. 데이터 품질의 차원을 나누는 문제는 데이터 품질을 향상시키는 문제뿐만 아니라 측정하는 문제에도 영향을 줄 수 있다. 예를 들면, 품질의 접근가능성을 측정함으로써 의사결정자가 이 데이터를 잘 이용하고 있는가 하는 문제를 파악할 수가 있기 때문이다.



<그림 4> 데이터 품질의 차원

4. 데이터 웨어하우스 환경에서 데이터 품질향상

4.1 데이터 품질향상을 위한 프레임워크 개발

데이터 웨어하우스 환경하에서 고려할 수 있는 데이터 품질 향상 요소로는 데이터 품질의 현재수준, 적절한 의사결정 과정을 위해 요구되는 품질의 수준, 데이터 품질을 강화하기 위해 설계된 프로젝트를 수행함으로써 얻게될 품질의 수준, 우선권을 가진 기업활동, 데이터 품질 향상 비용 등이다 [Ballou and Tayi, 1999].

이러한 데이터 품질 향상 요인들을 고려한 품질향상 프레임워크의 구성요소는 다음과 같다. 첫째, 품질 향상을 위한 노력으로 데이터 웨어하우스 관리자는 먼저 데이터 웨어하우스가 어떤 기업활동을 지원해야 하는지 의사결정자들에게 확인해야 한다. 둘째, 목표로 정해진 기업 활동을 지원하기 위해 요구되는 데이터세트(data sets)를 확인해야 한다. 여기서 데이터세트란 '거래파일' 혹은 '외부데이터의 집합'과 같이 명백하게 구별이 되는 데이터의 집합을 말한다. 셋째, 데이터 품질의 차원(data quality dimensions)은 데이터세트가 가지고 있는 현재 또는 미래의 잠재적인 문제들을 확인하는데 도움을 준다.

앞에서 검토한 구성요소를 토대로 데이터 웨어하우스 환경에서 데이터 품질의 향상을 위한 개념적 프레임워크를 제안하면 <그림 5>와 같다. 이 프레임워크는 기업활동, 데이터세트, 데이터 품질차원 등으로 구성된다. 본 연구에서 제안한 프레임워크는 특정한 기업활동과 특정한 데이터세트를 대상으로 구성하였으나, 데이터 웨어하우스 환경에서 데이터 품질향상을 위한 프레임워크로 일반화할 수도 있을 것이다.

본 연구에서 제안한 3차원의 데이터 웨어하우스 데이터 품질향상을 위한 프레임워크는 첫째, 기업의 목적에 적합한 품질향상 활동을 지원하고, 생산계획, 판매추적 등과 같은 다양한 기업활동을 가장 잘 지원할 수 있는 데이터 품질향상 활동이 이루어지도록 고려하였으며, 둘째, 웨어하우스 관리자의 데이터 품질향상 활동을 지원하기 위해서는 품질차원이나 데이터세트 등과 같은 품질향상에 필요한 다양한 문제들을 관리자가 인식할 수 있도록 구성하였다. 셋째, 데이터 웨어하우스 환경에서 데이터 품질 향상을 위한 총괄적이고 체계적인 방안을 제공할 수 있는 프레임워크를 개발하였다. 이러한 3차원의 데이터 품질 향상 프레임워크에 기초하여 데이터 품질의 향상을 위한 프로젝트 결정 메트릭스를 나타내면 <그림 6>과 같다.

이 메트릭스의 특징은 데이터 품질의 향상을 위한 9가지의 프로젝트가 수행될 수 있다는 것을 보여주고 있다. 이러한 프로젝트의 수행을 통하여 데이터 품질 향상의 실제를 확인 할 수 있다. 앞에서 제안한 품질향상 프레임워크를 지수표기법으로 나타내면 다음과 같다. 이러한 지수는 품질차원, 데이터세트, 기업활동, 품질향상 프로젝트 등 4가지 측면으로 구성된 품질향상 메트릭스의 12가지 구성요소를 나타내고 있다.

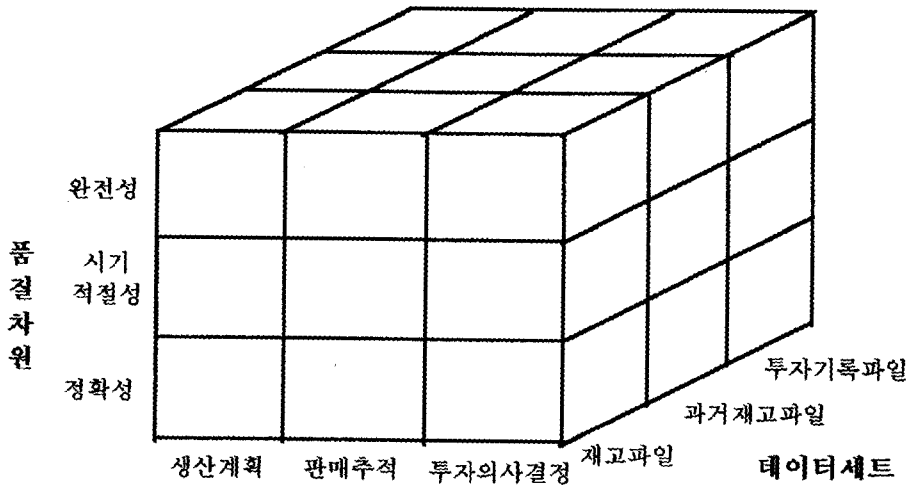
I : 데이터 웨어하우스가 지원하는 기업활동 지수
J : 데이터세트 지수
K : 데이터 품질의 차원 지수
L : 데이터 품질 향상 프로젝트 지수

앞에서 제안한 3차원의 데이터 품질향상 프레임워크에 근거하여 고안된 품질향상 프로젝트 결정 메트릭스를 설명하면 다음과 같다. 첫째, 데이터 웨어하우스가 생산계획(I=1), 판매활동(I=2), 투자의사결정(I=3) 등과 같은 3가지 기업활동을 지원한다. 이러한 활동들은 재고파일(J=1), 과거판매활동파일(J=2), 투자기록파일(J=3) 등과 같은 3가지 데이터세트에 의해 지원된다. 품질차원의 관점에서 재고파일은 정확성(K=1)이 다소 부족하며, 시기적절성(K=2)도 떨어진다. 과거판매활동파일은 정확하지만(K=1), 과거의 자료로서 시기적절성이 떨어진다(K=2). 투자기록파일은 완전성(K=3)을 보장할 수 없다.

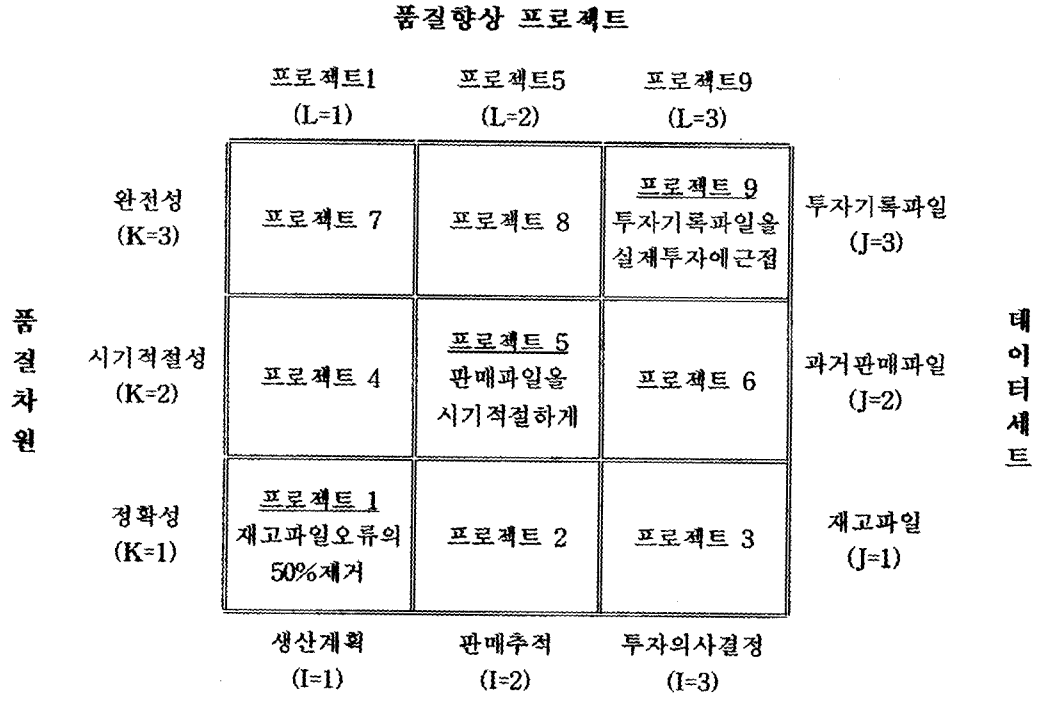
이러한 파일들의 품질을 향상시키기 위한 품질향상 프로젝트 결정 메트릭스에서 3가지 품질향상 프로젝트를 생각할 수 있다. (1) 재고파일의 오류의 50%를 제거하는 프로젝트(L=1), (2) 약간의 정확성이 결여된다 하더라도 과거판매활동파일을 좀 더 시기적절하게 만드는 프로젝트(L=2), (3) 투자기록파일이 실제 투자와 거의 일치하도록 완전하게 만드는 프로젝트(L=3)를 고려할 수 있다. 물론 한정된 자원으로 인해 세 프로젝트를 모두 수행할 수는 없다.

4.2 데이터 품질향상을 위한 프로젝트에서 고려해야할 요인

(1) 데이터의 현재 품질수준 ; CQ(J, K) : 각 데이터세트의 현재 품질은 각각의 데이터 품질 차원 K로 평가된다. 위의 시나리오에 따르면, 재고파일은 정확성 CQ(J=1, K=1)과 시기적절성 CQ(J=1, K=2)이 평가되어야 한다.



<그림 5> 데이터웨어하우스에서 데이터 품질 향상 프레임워크



<그림 6> 품질향상 프로젝트 결정 매트릭스

(2) 의사결정에 요구되는 품질 ; RQ(I,J,K) : 기업활동을 수행하는데 요구되는 품질의 수준으로 데이터 세트 J를 사용하고, 차원 K에 의해 좌우된다. 데이터 웨어하우스가 저장하고 있는 데이터는 여러 가지 목적을 위해 사용되므로, 데이터에 대한 품질 요구는 매우 다양할 수 있다.

(3) 예견된 품질 ; AQ(J,K;L) : 이 요인은 프로젝트

L을 수행함으로써 얻게 될 차원 K에서의 데이터세트 J의 품질 수준을 나타낸다. 데이터 품질 담당자는 품질에 영향을 주는 다양한 프로젝트를 수행할 수 있으며, 특별한 차원의 품질을 향상시키기 위한 노력이 다른 차원의 품질을 떨어지게 할 수도 있다. 위의 시나리오에서 프로젝트 L=2(과거판매활동파일 J=2를 좀 더 시기적절하게 만드는 프로젝트)가 수행되면, 시기적절성 차원(K=2)은 향상되나 정확성(K=1)은 손해를 보는 경우가 한 예가 될 것이다.

주의할 것은, 데이터 품질 향상 프로젝트가 데이터 품질을 향상시키려는 것이라 할지라도 반드시 그것이 최선이 되는 것은 아니다. 예를 들어, 의사결정에 어떤 수준의 데이터 품질을 요구하는데 현재의 품질 수준이 더 높다면 자원을 절약한다는 측면에서 품질을 더 낮추는 경우도 있을 수 있다는 것이다.

위의 3가지 요인(CQ, RQ, AQ)은 0에서 1까지의 범위를 갖는데, 가장 나쁜 경우는 0이 되고, 가장 좋은 경우는 1이 된다. 예를 들어, 재고파일에서 CQ(1,1)=0.6(재고파일 데이터세트의 60%가 정확하다.)인 경우를 가정해보자. 프로젝트 1은 오류의 50%를 제거하는 것이므로 AQ(1,1;1)=0.8이 될 것이다.

이 외에 고려해야할 요인들은 다음과 같다.

(4) 기업활동의 우선권; Weight(I) : 어떤 기업활동이 다른 활동보다 더 중요한 경우가 있다. 즉 모든 요인들이 동등할 때, 높은 우선권을 가진 활동이 낮은 우선권을 가진 활동에 포함된 데이터세트보다 더 선호되어야 한다. Weight(I)는 0보다 크고, 1보다 적어야 하며, 각 Weight(I)들의 합은 1이 되어야 한다.

(5) 데이터 품질 향상 비용; Cost(L) : 프로젝트를 수행할 때 드는 비용을 나타내며 자금, 인력, 시간과 같은 제약요소를 고려해야 한다. Cost(L)과 Weight(I) 사이에는 역관계가 존재할 수 있어서, 비용이 많이 들지만, Weight(I)가 큰 활동을 지원하는 프로젝트를 수행할 것인가, 아니면 중간정도의 중요성을 가졌지만, 비용이 적게 드는 활동을 지원하는 프로젝트를 수행할 것인가하는 선택의 문제가 생길 수 있다. Cost(L)은 어떤 기간동안 프로젝트 L을 수행하는데 든 전체비용을 나타내야 한다. 실제 프로젝트 작업 비용뿐만 아니라 프로젝트에 의해 요구되는 데이터 품질과 관련된 행동들을 수행하는데 쓰여진 모든 비용들이 포함되어야 한다.

(6) Value(L) : 데이터 품질 향상 프로젝트를 수행함으로써 얻을 수 있는 가치를 말한다.

(7) Utility(I,J,K;L) : 프로젝트 L이 수행될 때 데이터세트와 데이터 품질의 변화를 나타낸다. 프로젝트 L이 수행되었으나, 데이터세트가 영향을 받지 않는다면 Utility(I,J,K;L)=0이 되고, positive(프로젝트가 품질을 향상시키는 경우)하거나, negative(프로젝트가 품질을 악화시키는 경우)할 수도 있다. U(I,J,K;L) 값은 3×3×3×3=81 가지의 Utility(I,J,K,L) 값을 도출할 수 있다. 그러나 본 연구에서 제안한 위의 조건에서는 다음과 같은 6가지 Utility(I,J,K,L) 값을 고려할 수 있다.

이와 같이 U(I,J,K;L)의 값을 평가하는 이유는 데이터 웨어하우스 경영자들이 데이터세트와 데이터 품질의 차원이 특정한 프로젝트에 의해 어떤 영향을 받고, 그러한 영향의 이익의 상대적인 양이 얼마인지를 알지 못하면, 어떤 프로젝트가 수행이 되어야 하는지에 대한 명확한 의사결정을 할 수가 없기 때문이다.

U(1,1,1;1) = U(생산계획, 재고파일, 정확성; 재고파일 프로젝트)
U(1,2,1;2) = U(생산계획, 과거판매활동파일, 정확성; 과거판매활동파일 프로젝트)
U(1,2,2;2) = U(생산계획, 과거판매활동파일, 시기적절성; 과거판매활동파일 프로젝트)
U(2,2,1;2) = U(판매활동, 과거판매활동파일, 정확성; 과거판매활동파일 프로젝트)
U(2,2,2;2) = U(판매활동, 과거판매활동파일, 시기적절성; 과거판매활동파일 프로젝트)
U(3,3,3;3) = U(투자의사결정, 과거투자기록파일, 완전성; 과거투자기록파일 프로젝트)
Format : U(I:기업활동, J:데이터세트, K:데이터 품질 차원, L:품질 향상 프로젝트

Ballou와 Tayi[1999]은 정수형 프로그래밍 모형(integer programming model)을 제안하였다. 그들이 개발한 모형은 웨어하우스 데이터의 이용성을 최대화하고, 다양한 상충작용(trade-offs)을 체계적으로 통합하며, 현재는 이용할 수 없는 데이터를 획득하거나, 어떤 경우에 저장된 데이터의 양을 감소시킴으로써 이득을 올릴 수 있도록 하는 데이터 품질 프로젝트를 확인할 수 있다.

이제 데이터 품질향상 프로젝트의 사례 분석을 통하여 프로젝트의 중요도 우선순위를 분석하면 다음과 같다. 기업 A는 컴퓨터를 외주 또는 자체 생산하고, 판매에 중점을 두고 있는 회사이다. 다른 조건은 앞의 내용과 일치하며, Weight(I), Cost(l), U(I,J,K;L)은 다음과 같다.

이와 같은 조건을 감안하여 Ballou와 Tayi[1999]

Weight(I) : Weight(1)=0.3 Weight(2)=0.5 Weight(3)=0.2

Cost(L) : Cost(1)=5천만원 Cost(2)=1억5천만원 Cost(3)=4천만원

U(I,J,K;L) : U(1,1,1;1)=0.8 U(1,2,1;2)=0.2 U(1,2,2;2)=0.3
U(2,2,1;2)=0.1 U(2,2,2;2)=0.1 U(3,3,3;3)=0.9

Format : U(I:기업활동, J:데이터세트, K:데이터 품질 차원, L:품질 향상 프로젝트)

$$\text{Value(L)} = \sum_{All} Weight(I) \sum_{All} \sum_{All} Utility(I,J,K;L)$$ [Ballou & Tayi, 1999]

Value(1) = Weight(1) × U(1,1,1;1) = 0.3 × 0.8 = 0.24
Value(2) = Weight(1) × U(1,2,1;2) + Weight(1) × U(1,2,2;2) + Weight(2) × U(2,2,1;2) Weight(2) × U(2,2,2;2)
= 0.3 × 0.2 + 0.3 × 0.3 + 0.5 × 0.1 + 0.5 × 0.1 = 0.25
Value(3) = Weight(3) × U(3,3,3;3) = 0.2 × 0.9 = 0.18

가 제안한 정수형 프로그래밍 모형을 이용하여 3가지의 데이터 품질 향상 프로젝트를 수행함으로써 얻을 수 있는 가치 Value(L)을 계산하면 Value(1)=0.24, Value(2)=0.25, Value(3)=0.18 과 같다. Value(L) 값은 프로젝트2를 수행할 때 가장 큰 것으로 나타났다. 그러나, 자원이 제한되어 있는 상황에서 비용이 많이 드는 프로젝트 2를 수행하는 것은 기업 A에게 재무적 위험을 가져다 줄 수 있다. 따라서 비용 문제와 Value(L) 값의 가중치를 잘 고려한다면, 프로젝트 1을 수행하는 것이 기업에 보다 큰 이득을 줄 수 있을 것으로 본다.

5. 결 론

본 연구는 문헌연구를 통하여 데이터 웨어하우스 환경에서 데이터 품질의 향상을 위한 개념적 프레임워크를 개발하였다. 본 연구에서 제안한 데이터 웨어하우스 데이터 품질향상을 위한 프레임워크는 다음과 같은 역할을 수행할 수 있을 것으로 본다. 첫째, 기업의 목적에 적합한 품질향상 활동을 지원하고, 생산계획, 판매추적 등과 같은 다양한 기업활동을 가장 잘 지원할 수 있는 데이터 품질향상 활동이 이루어지도록 고려하였다. 둘째, 웨어하우스 관리자의 데이터 품질향상 활동을 지원하기 위해서는 품질차원이나 데이터세트 등과 같은 품질향상에 필요한 다양한 문제들을 관리자가 인식할 수 있도록 구성하였다. 셋째, 데이터 웨어하우스 환경에서 데이터 품질 향상을 위한 총괄적이고 체계적인 방안을 제공할 수 있는 프레임워크를 개발하였다.

데이터 웨어하우스 데이터는 다양한 요구를 가진 사용자들에 의해 접근되기 때문에 데이터 품질 활동은 기업 활동을 가장 잘 지원할 수 있도록 균형적으로 이루어져야 할 것이다. 본 연구에서 개발한 데이터 품질향상 프레임워크에 근거하여, 데이터 웨어하우스 환경하에서 데이터 품질향상을 위한, 데이터 웨어하우스 운영기업과 관리자에게 품질향상 지침을 제안하면 다음과 같다.

첫째, 데이터 웨어하우스 관리자는 데이터 품질향상을 위하여 데이터 웨어하우스가 지원할 수 있는 기업활동들을 구체적으로 결정한다.

둘째, 기업활동을 지원하는데 요구되는 데이터의 모든 세트를 정의한다.

셋째, 각각의 적절한 데이터 품질 차원에서 각 데이터세트의 품질을 평가한다.

넷째, 데이터 품질을 향상하기 위해 수행될 수 있는 잠재적인 프로젝트의 세트를 정의한다.

다섯째, 데이터 품질 차원에 따라 데이터 세트의 품질에 프로젝트들이 어떤 영향을 주었는지를 평가한다.

여섯째, 각각의 프로젝트, 데이터세트, 적절한 데이터 품질 차원에 대하여, 특정한 프로젝트가 수행될 때 이용성에서의 변화를 결정한다.

<참 고 문 헌>

1. 유사라, "데이터베이스 정보 품질 평가의 메타분석", 정보관리 학회지, 제16권 제1호, 1999, pp. 157-173.
2. 이국희, "온라인 데이터베이스의 품질평가", 데이터베이스 심포지움 및 학술대회 논문집, 1995년 11월, pp. 171-183.
3. 이병수, 이상락, 장근, "데이터 웨어하우스 구축 방법론에 대한 연구", 한국산업정보학회 논문지, 제4권 제2호, 1999, 6월, pp. 23-31.

4. 이재식, 전용준, "데이터 웨어하우스 설계를 위한 개념적 모델링 접근 방법", 데이터베이스 심포지움 및 학술대회, 1996년 11월, pp. 217-228.
5. 장동인, "데이터 웨어하우스의 최신동향", Oracle Magazine, 1997, pp. 44-78.
6. 전태형, 최재영, 최기준, "데이터 웨어하우스의 차원결정을 위한 요인에 관한 연구", 한국경영정보학회 국제학술대회논문집, 1998, 11월, pp. 455-464.
7. 정명호, 차원계층의 불균형에 의한 데이터 웨어하우스 검색의 이상현상에 대한 연구, 아주대학교 대학원 경영정보학과 석사학위논문, 1998.
8. 정희원, 데이터 웨어하우스를 고려한 효율적인 인트라넷 구축 방안에 관한 연구, 한양대학교 산업대학원 석사학위논문, 1997.
9. 조재희, "데이터 웨어하우징 기술 응용 그리고 미래", 한국경영정보학회 추계학술대회 논문집, 1996년 12월, pp. 29-40.
10. 조재희, "데이터 웨어하우징과 기업정보분석환경 리엔지니어링에 관한 연구", 한국경영정보학회 춘계학술대회 논문집, 1997, 6월, pp. 101-110.
11. 편집자, "데이터베이스 품질평가", 데이터베이스 월드, 1996년 4월, pp. 49-55.
12. 편집자, "데이터베이스 품질평가II", 데이터베이스 월드, 1996년 5월, pp. 69-80.
13. Arun, S., Varghose, S. J., Industrial-Strength Data Warehousing, *Communications of the ACM*, Vol. 41, No. 9, September 1998, pp. 29-31.
14. Ballou, D. P., and Pazer, H., Modeling Data and Process Quality in Multi-input, Multi-output Information Systems, *Management Science*, Vol. 31, No. 2, February, 1985, pp. 150-162.
15. Ballou, D. P., And Pazer, H, Designing Information Systems to Optimize the Accuracy-timeliness Trade-off, *Information Systems Research*, Vol. 6, No. 1, March 1995, pp. 51-72.
16. Ballou, D. P., and Tayi, G. K., Methodology for Allocating Resources for Data Quality Enhancement, *Communications of the ACM*, Vol. 32, No. 3, March 1989, pp. 320-329.
17. Ballou, D. P., and Tayi, G. K., Enhancing Data Quality in Data Warehouse Environments, *Communications of the ACM*, Vol. 42, No. 1, January 1999, pp. 73-78.
18. Ballou, D. P., and Tayi, G. K., Examining Data Quality, *Communications of the ACM*, Vol. 41, No.2, February 1998, pp. 54-57.
19. Bontempo, C., Zagelow, G., The IBM Data Warehouse Architecture, *Communications of the ACM*, Vol. 41, No. 9, September 1998, pp. 38-48.
20. Bort, J., The Wiser, Gentler Data Warehouse, http://www.sunworld.com/swol-05-1997/swol-05-datawarehouse.html, March 1997.
21. Celko, J., and McDonald, J., Don't Warehouse Dirty Data, *Datamation*, Vol. 41, No. 19, October, 1995, pp. 42-53.
22. Cushing, B., A Mathematical Approach to the Analysis and Design of Internal Control Systems, *Accounting Review*, Vol. 49, No. 1, January 1974, pp. 24-41.
23. David, K., Pamayya, K., Pema, P., and James, P., Assessing Data Quality in Information Accounting Systems, *Communications of ACM*, Vol. 41, No. 2, February 1998. pp. 72-78.
24. Diane, M. S., Yang, W. L., and Richard, Y. W., Data Quality in Context, *Communications of the ACM*, Vol. 40, No. 5, May 1997, pp. 103-110.
25. Gardner, S. R., Building the Data Warehouse, *Communications of the ACM*, Vol. 41, No. 9, September 1998, pp. 52-60.
26. Glassey, K., Seducing the End User, *Communications of the ACM*, Vol. 41, No .9, September 1998, pp. 62-69.
27. Ken, O., Data Quality Systems Theory, *Communications of the ACM*, Vol. 41, No. 2, February 1998. pp. 66-71.
28. Laudon, K., Data Quality and Due Process in Large Interorganizational Record Systems, *Communications of the ACM*, Vol. 29, No. 1, January 1986, pp. 4-18.
29. Morey, R. C., Estimating and Improving the Quality of Information in an MIS, *Communications of the ACM*, Vol. 25, No. 5, May 1982, pp. 337-342.
30. Redman, T., *Data Quality : Management and Technology*, Bamtam Books, New York, 1992.
31. Richard, Y. W., A Product Perspective on Total Data Quality Management, *Communications of the ACM*, Vol. 41, No. 2, February 1998, pp. 58-65.
32. Sutter, J. R., Project-Based Warehouses, *Communications of the ACM*, Vol. 41, No. 9, September 1998, pp. 49-51.
33. Thomas, C. R., The Impact of Poor Data Quality on the Typical Enterprise, *Communications of the ACM*, Vol. 41, No. 2, February 1998, pp. 79-82.
34. Wang, R. Y. and Strong, D. M., Beyond Accuray: What Data Quality Means to Data Consumers, *Journal of Management Information Systems*, Vol. 12, No. 4, Spring 1996, pp. 623-640.
35. Watson, H. J., Haley, H. J., Managerial Consideration, *Communications of the ACM*, Vol. 41, No. 9, September 1998, pp. 32-37.
36. Yair, W., and Richard, Y. W., Anchoring Data Quality Dimensions in Ontological Foundations, *Communications of the ACM*, Vol. 39, No. 11, November 1996, pp. 86-95.