데이터웨어하우스 구축 방법론에 대한 연구

(A Study of Impementaton Methodology for Data Warehouse)

이 병 수* 이 상 락** 장 근*** 윤 주 용****

(Lee Byong Soo Lee Sang Rak Chang Keun Yoon Ju Yong)

< 국문요약 >

기업들은 정보시스템을 활용하여 방대한 데이터를 신속하고 정확하게 처리를 함으로서 생산성을 높일 수 있는 기회를 갖게 되었다. 그러나 의사결정을 지원하는데 있어 정보시스템 기능의 한계가 자주 제기 되었다. 이에 따라 의사결정 지원시스템(DSS)의 중심을 이루는 수리적 모델의 응용 이외에 업무적 분석을 지원하기 위한 시스템으로서 데이터웨어하우스의 필요성이 제시되고 있다. 이를 위해 본 연구에서는 데이터웨어하우스의 개발 방법론과 관련한 문제 논의, 특히 개발과정과 관련된 다양한 모형을 소개하고 데이터웨어하우스의 성공을 위한 핵심사항과의 관련성을 고찰하였다.

<Abstract>

Using Information systems to process massive data, quickly and exactly, organizations have chances to enhance their performance. The limitations of IS function to support decision-making, however, have been frequently mentioned. In this context, in addition to traditional mathematical model that is a kernel of DSS, the needs for Data Warehouse which is a system supporting business process analysis are emerging. In this study, for those needs, first, we introduce issues of implementation methodology for D/W, especially various models relating development process. Then, we investigate correlation between these models and key factors for success of D/W.

* 시립 인천대학교 전자계산학과 교수

** 시립 인천대학교 전자계산학과 조교수

*** 시립 인천대학교 전자계산학과 강사

**** 한국 도로공사 전산부 부장

1. 서 론

오늘날 기업들은 정보시스템을 활용하여 방대한 데이터를 신속하고 정확하게 처리를 함으로서 생산성을 높일 수 있었다. 특히 단순 반복적(routine) 혹은 정형 적인(structured) 업무에 적용은 전통적인 데이터처리(data processing) 영역이라 할 수 있다. 그러나 의사결정을 지원하는데 있어 정보시스템 기능의 미흡한 점이 자주 언급되어왔다. 이에 따라 의사결정 지원시스템(DDS)의 중심을 이루는 수리적 모델의 응용 이외에 업무적 분석을 지원하기 위한 시스템으로서 데이터웨어하우스의 필요성이 제기된 것이다. 이러한 개념은 새로이 나타난 것은 아니지만 그것을 구현할 수 있는 기술적 한계는 최근에 극복 가능하게 되었다고 할 수 있다.

모든 시스템은 목적을 가지고 있다. 데이터웨어하우스는 특정한 업무영역 또는 기능영역(function)을 지원하는 기존의 시스템과는 다른 적용범위(domain)를 갖는다. 또한 설계의 문제에서조차 사용자를 중심으로 해야 한다는 특징을 갖는다. 즉, 기존

의 정보시스템의 개발방법론에서는 수용하기 어려운 특성을 갖고 있는 것이다.

따라서 본 연구에서는 데이터웨어하우스의 개발 방법론과 관련한 무제를 논의하고자 한다. 특히 개발 과정과 관련된 다양한 모형을 소개하고 데이터웨어 하우스의 성공을 위한 핵심 사항과의 관련성을 고찰 해보고자 한다.

2. 데이터웨어하우스의 개관

2.1 데이터웨어하우스의 정의

데이터웨어하우스는 '의사결정 지원을 위한 주제 지향적(subject-oriented), 통합적(integrated), 시간변이적(time-variant), 비휘발성(non-volatile)의 특성을 갖는 데이터 집합'이라고 할 수 있다 [Immon, 1993]. 이는 운영시스템 및 운영용 데이터베이스와의 차이를 중심으로 정의한 것으로 가장 일반적으로 인용되고 있다.

Poe는 '의사결정 지원시스템의 기초로 사용되는 읽기 전용의 데이터베이스'라고 정의한다. 한편, Kelly는 '기업 내에서 다영한 플랫폼 및 아키텍처 상에서 구현된 다양한 데이터 모델을 포함하고 있는 문제를 해결하기 위한 주제지향적 전사적 데이터베이스'라고 언급하고 있다.

이를 구체적으로 살펴보면, 데이터웨어하우스는 '직접 조회와 분석이 가능하도록 다양한 데이터베이스 혹은 정보원천(information sources)으로부터 채택된 데이터를 하나의 정보 저장소에 모으기 위한 아키텍처, 알고리즘, 툴들'을 포함하는 개념이라 할 수 있다. 또한 개발 과정을 포함하여 ' 데이터웨어하우징(data warehousing)'이라 언급되기도 한다.

2.2 데이터웨어하우스의 유용성

데이터웨어하우스의 사용에 따라 기대되는 혜택은 의사결정에 있어서 정보의 고유의(intrinsic) 질과 가치를 기초로 하는 것이 아니라, 접근 성을 기초로한 정보의 이용을 시간(stime span, and interval)과 사용자(user base) 측면에서 확장하는 것이다. 즉, 접근 가능한 데이터를 크게 늘림으로서 나타나는 것이다. 방대한 분량의 정보를 수집하는데 관리자가 소요하는 시간을 줄일 수 있으며, 환경에 관한 정보를 상시적으로 검토하는 기능을 포함시킬 수 있다. 또한 의사결정자의 추가적인 정보요구를 메타 데이터를 통하여 추적함으로서 효과적인 의사결정지원을 할 뿐 아니라 기업의 전략적 변화(혁신)를 추진하는 보조적인 역할을 할 수 있을 것이다.

2.3 데이터웨어하우스의 시스템적 특징

데이터웨어하우스는 기존의 정보시스템과는 상이한 시스템 적인 특징을 갖는다. 첫째, 요구분석이 사전적으로 결정되지 못한 상태에서 개발을 하여야 한다. 따라서 사용자의 요구사항을 파악하는 것이 반복적(iterative)으로 반복되어야 한다. 둘째, 처리하는 데이터가 운영시스템에서와는 다른 속성을 갖는다. 특히 시간이 하나의 키(key) 또는 필드(field)로 포함되어야 한다. 셋째, 시스템이 데이터의 처리(processing) 보다도 접근(access)과 적재(loading)를 중심으로 하는 것이다. 따라서 시스템에서의 입출력(input/output)이 처리시간 등을 고려한 설계에서 중요한 요인이 된다.

2.4 데이터베이스와 데이터웨어하우스

데이터웨어하우스는 사용자의 DB 종류(계층형, 관계형, 네트워크형)에 따라 상이한 솔루션을 가질 수 있다. 현재의 기술상태에서는 일반적으로 관계형 DB 솔루션이 적절하다고 한다. 조직의 업무특성에 따른 시간적 대응성의 필요 정도에 따라 온라인

분석업무의 요구를 결정하게 된다. 분석적 업무의 필요성 등에 따라 그 구축의 대상이 되는 데이터웨어하우스의 자체크기와 대상업무 범위를 결정하게 된다. 작은 규모 특히 특정 기능 부서를 지원하는 데이터웨어하우스를 데이터마트라고 부른다.

3. 데이터웨어하우스의 구축방법론

3.1 데이터웨어하우스의 설계를 위한 변수

1) 데이터모델

데이터모델이란 '데이터구조와 프로세스에 대하여 언어, 그림 숫자, 기타의 매체에 의해 추상화된 표현'이다. 데이터모델은 구조와 프로세스를 잘 이해할 수 있도록 해준다. 데이터모델은 데이터 설계에서 데이터 구조가 정확하도록 돕는다. 모델의 핵심은 효율적인 표현이다. 그러나 데이터 구조는 시간이 진행됨에 따라 확장된다. 이는 주제 영역에 대한 지식이 변화함에 따른 것이다. 데이터 모델링을 돕는 방법으로 잘 정립된 방법론과 CASE 툴이 있다.

데이터웨어하우스를 위한 데이터모델은 기업 데이터모델을 수정하여 사용한다. 기업 데이터모델은 기초적인 데이터로 구성되며, 운영시스템과 데이터웨어하우스의 구분 없이 적용할 수 있는 전체적인 데이터모델이다. 운영적 사용에서는 거의 변경되지 않지만, 데이터웨어하우스를 위해서는 상당부분 변경된다.

데이터웨어하우스의 데이터모델에서 핵심을 이루는 것은 바로 메타 데이터이다. 메타 데이터란 데이터에 관한 데이터이다. 데이터웨어하우스에서는 메타 데이터가 기존 정보시스템에서보다 더욱 중요하다. 메타 데이터가 데이터웨어하우스를 효과적으로 사용하는데 결정적이기 때문이다. 메타 데이터는 사용자인 분석가들이 정보를 검색하는 경로를 관리하는 방법을 제공한다. 메타 데이터는 ① 데이터의 구조, ② 원천 데이터, ③ 데이터의 변환, ④ 데이터모델, ⑤ 데이터모델과 데이터웨어하우스의 관계, ⑥ 추출에 관한 기록 등을 추적하는 정보를 가지고 있다. 운영적 환경에서 데이터웨어하우스 환경으로 사상(mapping) 하는 정보를 메타 데이터가 관리한다. 정보요구는 조직의 계층수준에 따라 그리고 기능에 따라 달라진다. 계층수준은 데이터의 요약정도(granularity)로 설계되고 기능은 메타 데이터로 설계될 수 있다.

2) 세밀화정도(Granularity)

세밀화정도는 요약의 수준과 집계의 수준으로 크게 나눌 수 있다. 요약이란 정보를 양적으로 감소시키고, 질적으로 풍부한 내용을 가진 정보로 만드는 것이다. 집계란 정보를 일정한 자원으로 합하여 총괄적인 정보로 만드는 것을 말한다.

① 요약의 수준 결정(Summarization)

요약된 데이터를 만들기 위한 세밀성의 정도를 정하기 위해서는 최종사용자에게 데이터를 보여주어야만 알 수 있다. 따라서 직접 사용할 시스템의 고객을 명확히 결정하여 이들의 요구를 통해서 파악하는 것이 필요하다. 따라서 빠른 프로토타이핑을 사용하는 것이 유용할 수 있다.

② 집계의 수준 결정(Aggregation)

데이터웨어하우스 설계에서 가장 중요한 측면은 집계수준이다. 집계수준은 데이터웨어하우스에서 단위 데이터에 적용되는 정밀도를 말한다. 특히 시간차원에서의 집계수준은 데이터의 분량결정에 매우 중요하다. 운영시스템에서 집계수준은 가장 낮은 수준의 데이터로 저장되는 것이 당연하다. 집계수준은 데이터웨어하우스에 상조하는 데이터의 양에 큰 영향을 미치며, 응답될 수 있는 조회의 유형에도 큰 영향을 미친다. 데이터웨어하우스에서 데이터의 양은 조회에서의 자세한 정도와 상충 하는 관계를 가지고 있다.

3) 구현 방법의 결정 : 온라인 분석의 필요성

데이터웨어하우스 구축 시에 온라인 프로세싱

을 어느 정도 수용할 것인가의 문제는 설계와 구축에서 유용성을 결정하는데 있어 매우 중요한 변수가 된다. 기업의 의사결정의 요구가 시간경쟁(time-bassed competition)의 중요성에 따라 달라질 것이지만, 분석업무의 적시성의 수준은 업무적 요구를 통한 별도의 분석을 필요로 한다. 결론적으로 OLAP의 필요성 문제는 데이터웨어하우스 구축의 유용성 문제와 보완적인 관계를 가진다. 다음의 <표 1>을 OLAP와 데이터웨어하우스의 보완적인 관계를 나타낸 것이다.

<표 1> OLAP와 데이터웨어하우스의 보완적인 관계

목 표		방대한 데이터지원	세밀 정도가 다른 다양한 수준의 지원	많은 데이터 분석요인(Analysis Factor)의 지원	많은 사이트에서 지원
적시성	사전 처리됨	DW	DW	DW	DW
	빠른 접근	OLAP	OLAP	OLAP	DW
	빠른 계산	OLAP	OLAP	OLAP	DW
정확성	정확한 원시 데이터	DW	DW	DW	DW
	연산의 풍부성	OLAP	OLAP	OLAP	OLAP
이해가능성	쉬운 인터페이스	OLAP	OLAP	OLAP	N/A
	유연한 보기 기능	OLAP	OLAP	OLAP	OLAP

4) 적재의 간격 결정

데이터웨어하우스에서 어느 정도의 시간간격을 두고 데이터를 상주시킬 것인가의 문제는 설계의 중요한 변수가 된다. 재 적재기간은 데이터모델의 수정과 의사결정 요구를 반영하는 주기를 이루게 된다. 따라서 재 적재는 의사결정 요구에 대한 시간적 주기와 밀접한 관련을 가지도록 설계되어야 한다. 이를 운영시스템과 독립적인 기능을 유지하는 적정한 기간으로 고려될 수도 있다. 특히 의사결정의 지원을 위해 최적화된 시스템으로 정착되는 것이 성공의 지표라 할 때 다양한 목적을 충족시키려는 시도는 실패 위험을 매우 크게 만든다. 또한 시간 간격은 데이터웨어하우스에 옮기기 전에 데이터를 안정시킬 기회를 제공한다. 이는 데이터의 품질을 높이는 역할을 수행한다.

이를 데이터의 주기성으로 파악할 수도 있다. 데이터의 주기성이란 운영환경에서 일어난 데이터의 변화가 데이터웨어하우스 환경에 반영되기까지의 시간 길이를 말한다. 예를 들어 24시간 의 간격이 있으면 데이터웨어하우스에서 운영프로세스를 수행하거나 운영프로세스에서 데이터웨어하우스 프로세스를 수행하려는 유혹을 미연에 방지할 수 있다.

운영환경과 데이터웨어하우스 환경을 결합시키는 정도가 강할수록 기술적으로 복잡해지고 비용이 많이 든다. 시간적 지연이 있는 경우 환경에의 특정한 규율을 적용 가능하게 한다.

3.2 데이터마트

1) 데이터마트와 데이터웨어하우스

데이터웨어하우스와 데이터마트는 개념이 명확히 구분되지 않을 정도로 혼용되어 사용되기도 한다. 보편적으로 전사적 의사결정의 지원과 부서별 지원이라는 접에서 구분되어진다. 특히 동시에 설계할 경우에 확장성의 문제와 관련하여 논의가 활발하다. 다음 <표 2> 는 데이터웨어하우스와 데이터마트를 구

<표 2> 데이터웨어하우스와 데이터마트의 비교

구 분	데이터웨어하우스	데이터마트
범 위	애플리케이션에 독립 중앙집중형, 공유됨 여러 업무에 걸침/ 기업 의도적으로 설계됨	특정의 애플리케이션 요구 여러 업무에 걸치거나, 부서, 사용자 영역 업무 - 프로세스 지향 중복 데이트를 가진 여러 개의 데이터베이스
데이터 관점	역사적 - 세부 데이터 약간의 요약 약간 비정규 화됨	세부적(일부 역사적) 요약됨 고도로 비정규 화됨
주 제	다양한 주제 영역	단일 주제 여러 개의 부분 주제 영역 운영적 원천의 스냅샷
데이터의 원천	많은 운영적, 외부 데이터	적음 운영적, 외부 데이터 OLTP 데이터베이스 스냅샷 '부트 레그' 데이터 추출
구현 시간대	첫단계에서 9~18개월 (2~3개의 주제 영역) 다 단계 구현	4~12개월
특 성	유연성 장기적/ 전략적 데이터 지향	제한적 단기적 수명/ 전술적 프로젝트 지향

분하는 가트너 그룹(Gartner Group, Inc.)이 제시한 기준이다.

2) 방대한 양

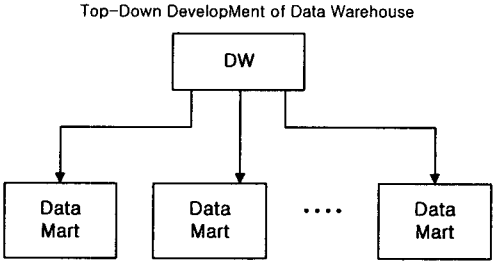
데이터웨어하우스 구축을 위해 대상이 되는 방대한 양의 데이터를 관리하도록 프로파일레코드(profile record)를 만드는 것이 효과적이다. 여기서 데이터웨어하우스가 다루어야 할 데이터의 양이 방대하거나 내용이 자주 바뀌는 경우에는 특별한 종류의 레코드를 필요로 한다.

여기서 다수의 운영데이터를 묶어주는 하나의 레코드를 프로파일 레코드 또는 집계 레코드(aggregate record)라고 부른다. 이것은 모든 데이터 구조에 들어가게 된다. 특히 양적인 관리에서는 프로파일 또는 종합 레코드를 만드는 것이 가장 중요한 기법이 된다. 반면에 이에 따른 데이터웨어하우스의 능력과 가능성이 손상되는 단점도 갖는다. 그러므로 DSS 분석가에게 세밀함이 얼마나 중요한지를 먼저 평가하도록 하여 프로파일 레코드를 설계하여야 한다.

프로파일 레코드를 주기적, 반복적으로 만든다면 매우 중요한 요소인 세밀함을 놓치는 오류를 방지할 수 있다. 데이터웨어하우스의 주기적인 반복활동이 짧고, 신속하면 변경에 있어 조율성을 높일 수 있다. 여기서 중요한 세밀성을 보장하는 또다른 방법으로는 세밀성이 다른 수준에서도 레코드를 만드는 것이다.

3) 하향식 설계(Top-down) 모델

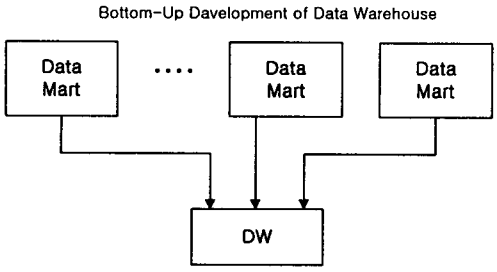
켄 오어는 데이터마트의 설계와 데이터웨어하우스의 설계를 연관하여 하향식 설계를 [그림 1]와 같이 정의한다.



[그림 1] 하향식 설계

4) 상향식 설계(Bottom-up)

또한 데이터마트의 설계와 데이터웨어하우스의 설계를 연관하여 상향식 설계를 [그림 2]와 같이 정의한다.



[그림 2] 상향식 설계

4. 결 론

데이터웨어하우스의 구축을 위한 문제영역은 첫째, 데이터모델의 필요성. 둘째, 모델에 데이터를 상주시키기 위한 전략. 셋째, 사용자가 유용한 정보를 조회하는 방법의 3가지 영역으로 크게 구분할 수 있다. 여기서는 데이터웨어하우스의 구축문제로서 ① 요구사항의 분석, ② 순환개발, ③ 데이터마트의 확장성 문제를 다루었고, 설계변소로는 세밀화 정도(Granularity), 온라인 질의의 필요성, 데이터 주기성을 다루었다. 이외에 의사결정모델, 분석 업무의 요구를 별도로 분석하고, 정보의 시간차원에 대한 보다 깊은 연구가 필요할 것이다.

데이터웨어하우스 설계 및 구축에서는 사용자의 요구사항의 분석, 개발기간, 외부 데이터의 고려, 다목적으로 설계하는 문제 등이 성패와 관련된다. 특히 기존의 사례 연구결과에서 데이터웨어하우스를 구축하는 기업의 특징으로 첫째, 경영관리를 정보중심으로 접근. 둘째, 매우 경쟁적이고 급변하는 시장환경. 셋째, 많고 다양한 고객기반. 넷째, 데이터가 저장된 시스템의 다양성. 다섯째, 동일한 데이터의 시스템간에 상이하게 표현됨. 여섯째, 포맷의 해석과 정돈의 어려움 등의 특징을 가진다. 따라서 데이터웨어하우스 구축에서 산업의 특성과 고객의 특성에 따라 성패에 미치는 영향이 클 수 있다.

참고 문헌

1. Widom J. "Research Problem in Data Warehousing," Proceedings of 4th International Conference and Knowledge Management (CIKM), Nov. 1995.

2. Immon, B. Building the Data Warehouse, New York: John Wiley & Sons, 1996.

3. Strange, K. "Can Data Marts Grow?, " CIO magazine, July 1, 1997.

4. Browne, M. Organizational Decision Making and Information, New York: Ablex Corp., 1993.

5. Thomsen, E. OLAP Sloutions, New York: John Wiley & Sons, 1997.

6. Norman, R. J. Object-Oriented Systems Analysis and Design, New Jersey: Prentice Hall, 1996.

7. Sprague, R. H. Jr. "A Framework for the

Development of Decision Support Systems," MIS Quarterly, Dec, 1980.

8. Epstein, B. J. The Multidimensional Value of Information, Michigan: UMI, 1980.

9. Drucker, P. F. " The Information Executives Truly Need," HBR, Jan-Feb, 1995.

10. John Ladley, " A Flexible Apporach to Developing a Data Warehouse," Data Warehouse - Practical Advice From the Experts, Prentice Hall, 1997, pp.100-119.

11. Ralph Kimball, Data Warehouse Architect, DBMS magazine, Dec., 1996.