經營論叢 第28輯
2007. 12. pp. 1~12

메타데이터 기반 데이터 통합 아키텍처

강 양 석[*1]
허 진 석[**2]
홍 순 구[***3]

<목 차>

Ⅰ. 서 론

Ⅱ. 선행 연구
1. 메타데이터의 정의 및 모델링
2. 메타데이터와 데이터웨어하우스
3. 데이터 품질이 기업에 미치는 영향
4. 메타데이터 관리의 현황 및 중요성
5. 메타데이터의 관리 방안
6. 메타데이터를 이용한 데이터 관리 방법

Ⅲ. 메타데이터 기반 데이터 통합 아키텍처
1. 통합 아키텍처의 정의
2. 통합 아키텍처의 제시

Ⅳ. 결 론

참고문헌

Ⅰ. 서 론

오늘날의 데이터 관리의 주요 현안은 실시간으로 획득되는 이질적인 대량의 데이터들을 어떻게 통합·처리 할 것인가에 관한 것이다. 이를 지원하기 위해 데이터베이스의 개념은 단일 부서의 정보 자원들을 관리하는 파일 시스템의 개념에서 전사적인 정보자원을 통합 관리하는 데이터웨어하우스의 개념으로 진화되었다.

데이터웨어하우스는 원시 데이터 계층, 데이터웨어하우스 계층, 클라이언트 계층으로 구성되며 데이터의 추출, 저장 및 조회의 기능을 수행한다. 운영 시스템은 조직 운영에 필요한 재고관리, 회계정보 및 영업 시스템 등과 같은 특화된 기능을 지원하지만 데이터웨어하우스

1) 동아대학교 경영정보과학부 석사과정
2) (주)이노베이티브데이터솔루션즈, 책임연구원
3) 동아대학교 경영정보과학부 부교수

는 고객, 제품 및 회계 등과 같은 주제를 중심으로 데이터를 구축한다.

데이터웨어하우스는 데이터의 재구성을 위해 원시 데이터 계층에서 획득된 서로 다른 데이터들을 이해할 수 있어야 하며 각각의 데이터들에 대한 이력을 유지해야 한다. 이러한 데이터웨어하우스의 데이터 통합 관리는 데이터웨어하우스 내의 메타데이터 관리를 통해 가능하다.

그러나 Eckerson(2004)의 연구에 따르면 전사적인 메타데이터(global metadata)의 관리는 기업의 지능화(business intelligence) 실현을 위한 중요한 요소임에도 불구하고 이를 위한 노력은 미비한 것으로 나타났다. 그 주된 이유가 데이터웨어하우스 내에 사용된 데이터를 위한 매트릭스(Metric)와 차원(Dimension)을 정의한다는 것은 매우 힘든 일이며, 정의된 메타데이터를 유지·보수하기 위한 방안이 많지 않다는 것과 인력을 통한 메타데이터의 규칙적인 관리가 비용과 시간 문제로 인해 잘 이루어지지 않는 다는 것이다.

본 연구에서는 데이터 관리의 문제점을 해결하기 위한 방안으로 메타데이터를 기반으로 한 데이터 통합 아키텍쳐를 제시하고자 한다. 이를 위해 현재의 메타데이터 관리가 어떠한 문제점을 가지고 있는지를 선행 연구를 통해 파악하고, 파악된 문제를 해결하기 위한 방안과 제시된 방안의 기술적 실현을 위한 전략으로써 메타데이터 관리 방안에 대한 프레임워크를 제시하고자 한다.

Ⅱ. 선행연구

2.1 메타데이터의 정의 및 모델링

메타데이터란 데이터에 대한 데이터라고 할 수 있다. 데이터 또는 데이터 셋(data set)을 효율적으로 접근하고 관리할 수 있도록 해 주는 데이터에 대한 정보를 총칭한다. 광의로는 데이터의 생성에 따른 기본적 내용, 질적 요소, 문서 구조, 구현 기법, 참조 정보, 데이터 내용에 관한 서술 정보, 데이터 접근, 획득, 배포, 활용에 관한 정보, 생성자, 관리자 정보 등을 광범위하게 기술하는 데이터 셋 또는 정보를 뜻하기도 한다.

어떤 대상 분야의 데이터에 대한 메타데이터를 구축하기 위해서는 메타데이터의 생성규칙, 사용 용어, 표기법과 같은 문법을 정의하고, 메타데이터의 구조 및 구성요소, 각 구성요소의 의미, 데이터 타입 및 데이터 영역 등을 정의해야 하는데 이와 같이 특정 분야에 대한 메타데이터의 정형적 모델을 정의하는 것을 메타데이터의 모델링이라고 한다.

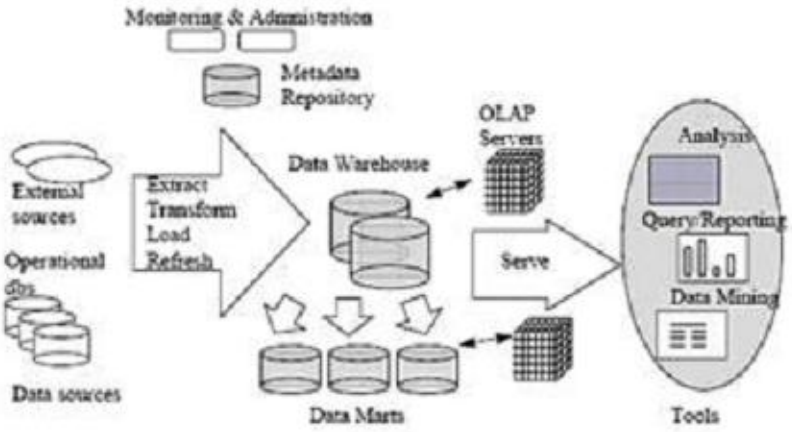
메타데이터의 모델링 시 고려할 사항으로는 계층형, 관계형, 객체형, 확장 관계형 등과 같

은 범용적이고 표준적인 데이터 모델을 채택하되 확장성, 이식성, 개방성 등이 가능한 데이터 모델이어야 하며, 대상 미디어별 특성 요소 및 응용 분야의 특성에 대한 메타데이터를 수용할 수 있어야 한다. 메타데이터의 저장, 관리를 위해 파일시스템 또는 DBMS 등의 하부구조를 고려해야 하며, 메타데이터 자체에 대한 GUI 기반의 브라우징 및 검색이 가능한 구조로 이루어져야 한다. 특히 이미 표준화된 참조 모델의 수용 또는 보완 적용 등도 고려할 수 있다.

2.2 메타데이터와 데이터웨어하우스

Chaudhuri(1997)의 연구에 따르면 데이터웨어하우스의 일반적인 구조는 <그림 1>과 같다. <그림 1>에 따르면 데이터웨어하우스는 다양한 운영 데이터베이스로부터 데이터를 추출(extracting) 할 수 있는 도구를 포함한다.

<그림 1> 데이터웨어하우스의 구조



[Chaudhuri, 1997, p.518]

<그림 1>에서 데이터웨어하우스와 데이터 마트의 데이터는 하나, 혹은 그 이상의 데이터웨어하우스 서버로 저장되며, 최종사용자 이전의 도구들에 의해 다양한 차원의 뷰(views)로 표현된다. 그리고 데이터웨어하우스는 로드 밸런싱(load balancing), 확장성(scalability) 및 적용성을 만족시켜야 하는데 이런 조건의 충족을 위해 메타데이터의 저장소는 데이터웨어하우스의 각 부분들과 일치하여야 하며 전체의 데이터웨어하우스는 일원화된 방법을 통해 관리되어야 한다.

데이터웨어하우스가 기업의 비즈니스 모델을 반영할 때, 메타데이터의 관리는 데이터 웨어하우징 아키텍쳐를 위한 필수 요소이며 이때 관리 메타데이터(Administrative metadata), 비즈니스 메타데이터(business metadata) 및 운영 메타데이터(operational metadata) 등의

다양한 종류의 서로 다른 메타데이터가 관리되어야 한다.

한편 메타데이터 저장소(repository)가 데이터웨어하우스 내의 모든 종류의 메타데이터를 저장 및 관리하기 위해 사용된다. 이 저장소는 데이터웨어하우스를 위한 도구들을 통해 메타데이터의 공유를 가능하게 한다. 상업적인 메타데이터 저장소에는 플래티넘 리파지토리(platinum repository)와 프리즘 디렉토리 메니저(prism directory manager) 등의 예가 있다.

2.3 데이터 품질이 기업에 미치는 영향

Ramchandra(2006)의 연구에 따르면 기업들은 데이터 품질 문제에 노출되어 있으나 이로 인한 직접적인 문제가 발생하기 이전의 기업들은 이런 문제를 인식하지 못하는 것으로 조사되었다. 저품질의 데이터가 기업에 미치는 영향에 대한 사례로 SIC (Standard Industry Classification) 데이터가 부정확할 경우 은행의 고객 분류의 오류로 인해 잘못된 추정의 원인이 되었다. 그리고 모 은행의 대출 시스템은 한 명의 고객이 서로 다른 십여개의 식별자(identifier)를 가지고 있었으며, 어떤 금융 기관의 경우 35개의 서로 다른 회계 코드(account status code)를 사용하는 것으로 나타났다. 이러한 저품질의 데이터가 기업에 미치는 손실은 미국의 경우 매년 미화 6000억 달러에 이르는 것으로 추정되었다.

2.4 메타데이터 관리의 현황 및 중요성

Eckerson(2004)의 연구에 따르면 전사적인 메타데이터(Global Meta Data)의 관리는 기업의 지능화(Business Intelligence) 실현을 위한 중요한 요소임에도 불구하고 이를 위한 노력은 미비한 것으로 나타났다. 데이터웨어하우스 내에 사용된 데이터를 위한 매트릭스(metric)와 차원(dimension)을 정의한다는 것은 힘든 일이며 기 정의된 메타데이터를 유지·보수하기 위한 방안은 많지 않다. 관리자들을 통한 메타데이터의 규칙적인 관리도 드문 것으로 조사되었다.

메타데이터는 실 데이터가 어떤 의미를 가지고 있는지 알기 위한 참조 데이터로 고객, 제품 및 공급자 리스트 등과 같은 주요 데이터들의 정보를 담고 있다. 일반적으로 이러한 데이터들은 다중 운영 시스템 내에 존재하며, 각각은 서로 다른 독특한 포맷(format)에 의해 관리된다. 시스템 간의 중복 데이터의 비율은 전체 데이터의 약 30%에서 50%를 차지하고 있으며 이는 기업의 효율성과 수익성을 저해 시키는 요인이다. 중복 데이터로 인해 중복 발송된 이메일은 메일 발송 비용을 증가시키며 같은 메일을 이중으로 받게 되는 고객의 불만을 야기시킬 수 있다. 생산 현장에서의 중복 레코드는 송장처리의 실수를 낳게 하며, 불안정한 생산 및 재고부족의 원인이 되기도 한다. 공급 사슬망에서의 부정확한 데이터는 기업, 공

급자, 분배자 간의 관계를 정확히 파악하지 못하게 하여 많게는 수백만 달러의 금전적 손실을 초래하기도 한다(Eckerson, 2004).

2.5 메타데이터의 관리 방안

Eckerson(2004)의 연구는 메타데이터를 관리하기 위한 기업의 3가지 접근 방안을 제시하고 있다. 첫째, 기업은 메타데이터의 표준화를 위해 전사적으로 단일 ERP 어플리케이션을 사용하여야 한다. 대부분의 기업은 시스템 변경을 위해서 레거시 시스템을 멈추지 않으며, 조직내 서로 다른 그룹에서 서로 다른 버전의 ERP 소프트웨어를 서로 다른 시각에서 업그레이드 시킴으로써 데이터 품질면에서 내부적인 결함을 발생 시킨다.

둘째, 기업은 표준화된 메타데이터 관리를 위해 데이터 저장을 위한 단일 어플리케이션을 설계하여야 한다. 예를 들어 기업은 공급자의 프로파일 관리를 위해 새로운 공급 관리 어플리케이션을 자사에 설치 할 수 있다. 이 어플리케이션은 정보를 사용자의 요청에 의해 다른 어플리케이션으로 분배한다. 이런 경우 어플리케이션의 관리자는 업데이트 및 변경과 관련된 이력을 포함한 데이터 관리에 책임이 있다.

마지막으로 기업이 데이터의 통합 관리를 위한 특별한 도구를 가지고 있지 않을 경우, 기업은 데이터 통합 허브(data integration hub)라 지칭되는 서버 엔진의 도입을 고려할 수 있다. 서버 엔진의 역할은 다양한 운영 시스템 간의 레퍼런스 데이터(메타데이터)의 분배, 표준화 및 동기화를 담당한다. 허브 엔진은 표준화된 레코드들의 사본을 유지하여 고객들의 레코드를 관리한다. 예를 들어, 마이크로소프트사는 내부적으로 고객의 데이터 관리를 안정화 및 표준화하기 위해 동기화 허브(synchronization hub)를 구현하였는데 이 허브는 현재 60여개의 서로 다른 시스템으로부터 들어오는 정보를 원천으로 약 1.8 테라바이트 분량의 고객 데이터로 구성되어 있다. 허브는 매칭 엔진(matching engine)을 통해 이중 데이터를 분별하고 유일한 식별자를 부여하며 특정 필드 값들을 표준화하여 저장한다.

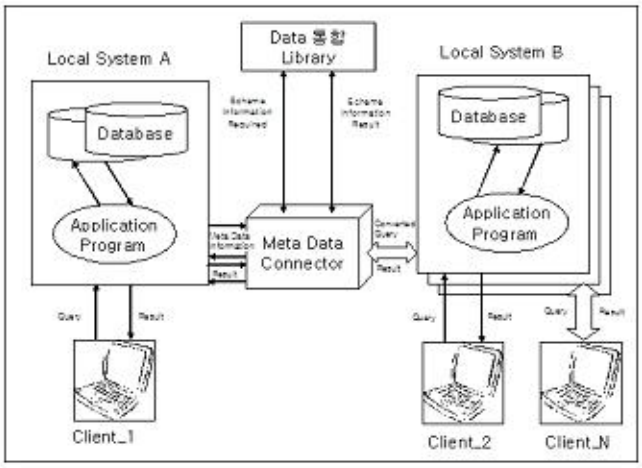
2.6 메타데이터를 이용한 데이터 관리 방법

Sciore(1998)는 데이터 통합 방법의 하나로 메타데이터를 이용하여 이 기종간의 데이터베이스간 의미상으로는 같지만 서로 다른 객체 이름과 특성을 가진 데이터를 공유할 수 있는 방법을 제시하였다.

이진수(2002)는 속성(attribute)을 통합하기 위해 별도의 메타데이터 표준안을 사용하지 않고, 데이터베이스 스키마의 의미와 인스턴스 값이 가지는 의미를 공유할 수 있는 시스템을 제안하였다. 데이터베이스 속성을 공유하기 위해 각 데이터 시스템에서 가지고 있는 속성의

특성 값을 통합된 데이터 라이브러리로 구축하고, 모든 시스템에서 구축된 라이브러리를 참조하는 방법을 이용하였다. 이때 각각의 데이터 시스템은 자신이 가지고 있는 속성 정보를 통합 라이브러리로 제공하고 다른 데이터 시스템이 가지고 있는 속성 정보를 참조 할 수 있게 된다. 설계된 통합 시스템의 모형은 <그림 2>와 같다.

<그림 2> 데이터 통합 구조



[이진수 외, 2002, p.47]

통합 시스템에서 질의에 대한 결과를 추출하는 과정은 다음 순서를 따른다.

1. 시스템 A의 클라이언트에서 응용 프로그램에 질의를 요청한다.
2. 질의를 요청받은 응용 프로그램은 자신의 데이터베이스에서 질의에 해당하는 응답을 임시 테이블_1에 저장한다.
3. 새롭게 생성된 임시 테이블_1의 속성에 관련된 정보를 메타데이터 변형자로 전달한다.
4. 메타데이터 변형자는 속성에 관련된 스키마 정보를 라이브러리를 통해 시스템 B의 속성으로 변형한다.
5. 시스템 B의 속성에 대한 정보를 이용해 새로운 변형질의를 도출한다.
6. 변형된 질의를 시스템 B로 전달한다.
7. 시스템 B에서는 변형된 질의를 받아 임시 테이블_2에 저장하여 시스템 A로 전달한다.
8. 시스템 A의 응용 프로그램은 시스템 B로부터 전달받은 임시테이블_2의 내용과 자신의 임시 테이블_1을 통합한다.
9. 통합된 결과 테이블_new를 클라이언트의 응답으로 전달한다.

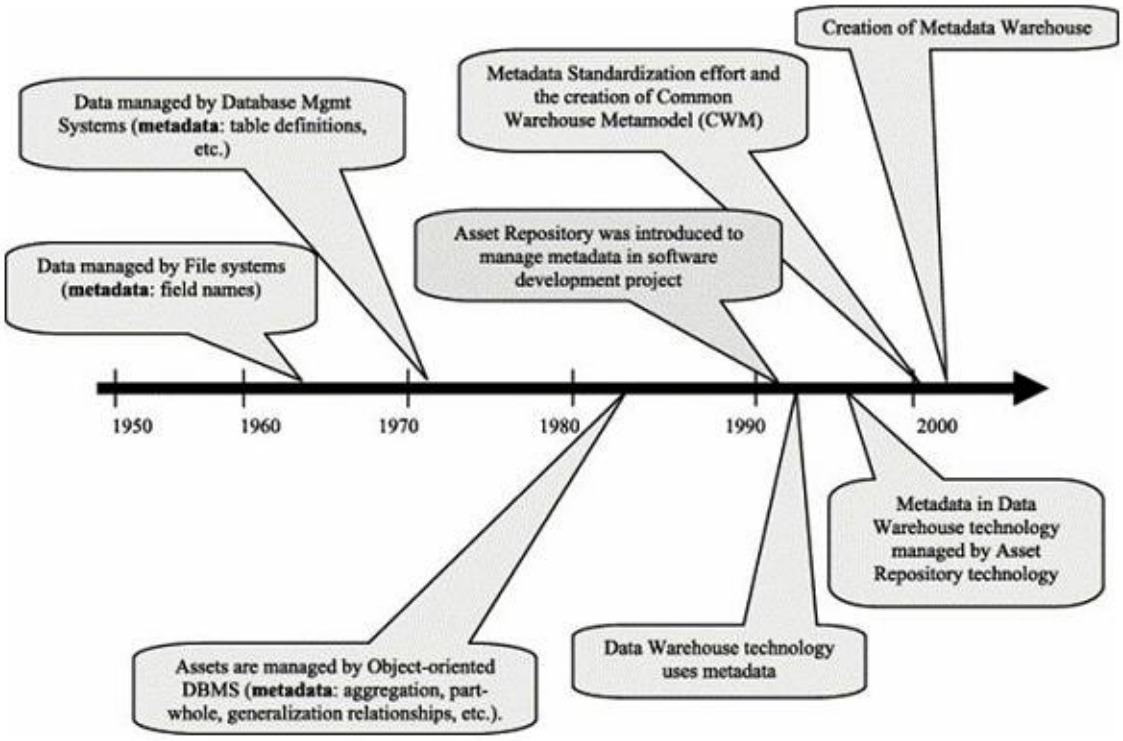
[그림 2]에서 보는 바와 같이 메타데이터 변형자는 시스템 A로부터의 질의를 시스템 B에

부합되는 질의로 변형하기 위해 데이터 통합 라이브러리를 참조하여 변형된 질의를 시스템 B로 전달하게 된다.

2.7 통시적 관점에서의 메타데이터 관리

Sen(2004)의 연구에 따르면 메타데이터의 관리는 전통적인 데이터베이스와 데이터 웨어하우스 관리 문제에 있어서 부차적인 것이었지만 오늘날의 기업들은 도구(tools) 및 데이터의 통합과 변화 관리(change management)를 위해 메타데이터를 필요로 한다고 언급하고 있다. 연구에 따르면 메타데이터 관리 도구(metadata repository)의 개념은 이미 사용되고 있으며, Sen(2004)은 거기에서 한걸음 더 나아가 'metadata warehouse'라는 새로운 개념의 메타데이터 관리 방법을 제시하였다. Sen(2004)의 연구에 의한 통시적 관점에서의 메타데이터 관리 방법의 진화는 <그림 3>과 같으며 메타데이터의 개념은 단순한 것에서 복잡한 것으로, 데이터 관리를 위한 중요도는 낮음에서 높음으로, 메타데이터와 원래 데이터의 성질의 차이는 유사한 수준에서 뚜렷한 차이를 보이는 수준으로 변화되었음을 알 수 있다.

<그림 3> 통시적 관점의 메타데이터 관리 방법의 변화



[Sen, 2004, p.154]

Sen(2004)은 메타데이터 웨어하우스 구축 방법을 다음과 같은 3가지로 설명하였다.

1. 제 1 접근법(데이터 웨어하우스의 구조를 이용)
2. 제 2 접근법(메타데이터 관리도구 기반 구조)
3. 제 3 접근법(통합 아키텍처)

Sen(2004)의 제 3 접근법에 의하면 최종 사용자는 다음과 같은 도구를 통해 메타데이터의 관리를 수행할 수 있다.

- 버전(version) 관리 도구
- 메타데이터 평가 도구
- 메타데이터 변경 관리 도구
- 의사 결정 지원 도구
- 웹 접근이 가능한 브라우징(browsing) 도구

본 연구에서는 Sen(2004)의 개념을 계승하여 메타데이터와 데이터 관리 방법들의 통합 수준이 아닌 메타데이터를 근간으로 하는 데이터 관리 방법에 대해 소개하고자 한다.

Ⅲ. 메타데이터 기반 데이터 통합 아키텍처

3.1 통합 아키텍처의 정의

선행연구에 따르면 데이터웨어하우스 내의 메타데이터에 대한 관리는 새로운 개념이 아니지만 데이터웨어하우스 내의 메타데이터와 그 외적인 요소들의 결합도가 어느 정도인가에 따라 메타데이터를 기반으로 한 데이터 통합 아키텍처 여부를 판단할 수 있다. 본 논문에서의 메타데이터를 기반으로 한 데이터 통합 아키텍처란 조직의 데이터 자원을 관리하기 위하여 데이터의 관리 구조를 메타데이터를 기반으로 하거나 또는 이를 적극적으로 활용하는 데이터 통합 구조로 정의한다.

3.2 통합 아키텍처의 제시

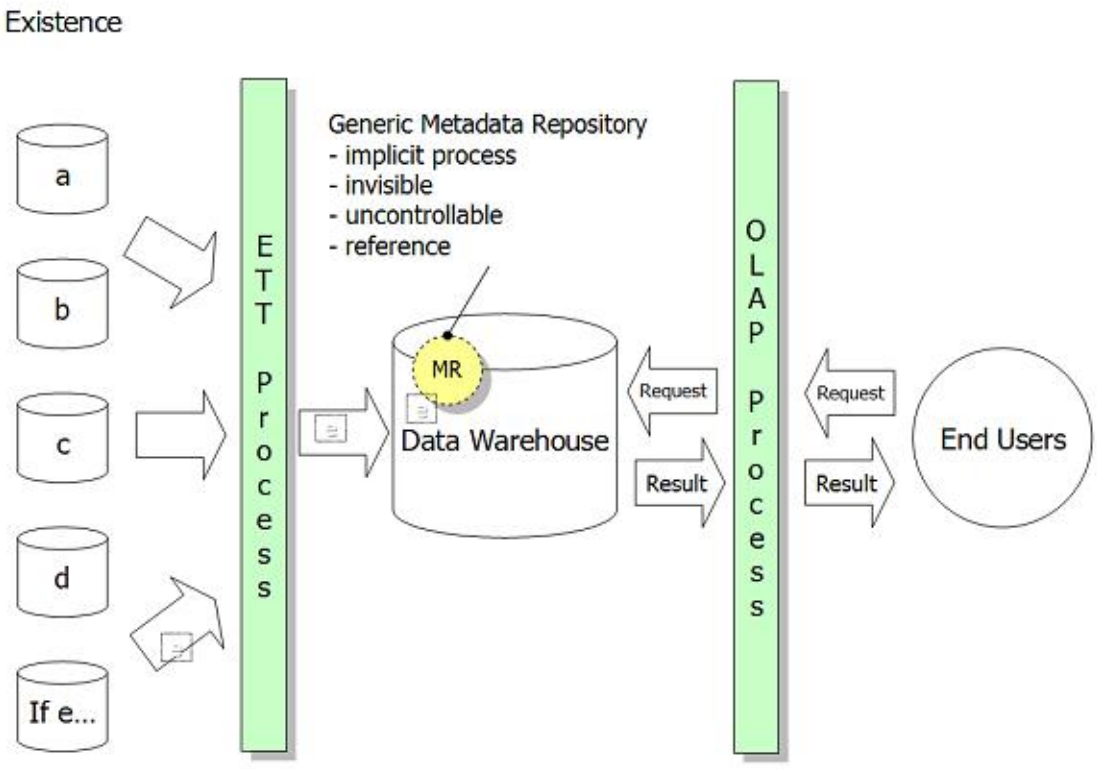
3.2.1 일반적인 데이터 관리 구조

<그림 4>에서 a, b, c, d, e는 각 부서로부터 발생된 비즈니스 메타데이터를 의미한다. 그리고 각 부서로부터 생성된 데이터들은 추출·전송·전환 과정을 거쳐 메타데이터 관리 도구와 데이터웨어하우스 내에 존재하게 된다.

이중에서 e는 설계부서에서 신제품 개발을 위해 새로이 생성한 비즈니스 메타데이터 정보이다. 설계부서의 경우 '고객'의 개념이 제품의 구매 및 판매 차원이 아니라는 점에서 다른

부서와 다소 상이하므로 고객이라는 일반적인 의미의 영문표기 대신 '설문이 수행된 개인'이라는 의미로 surveyed individual의 약어인 svyind 데이터 필드를 추가 하였다. 이 때 신규 발생된 비즈니스 메타데이터관련 정보인 e의 경우 데이터웨어하우스 내의 메타데이터 관리 도구(metadata repository)에 의해 갱신 및 유지되어야 하는데 선행 연구사례에서 언급된 바와 같이 여러 가지 원인에 의해 관리가 소홀한 실정이며 이런 경우 전사적인 데이터의 품질은 낮아지게 된다.

<그림 4> 일반적인 데이터웨어하우스와 메타데이터와의 관계



※ 영문자는 데이터 자원들 내의 메타데이터 관련 데이터

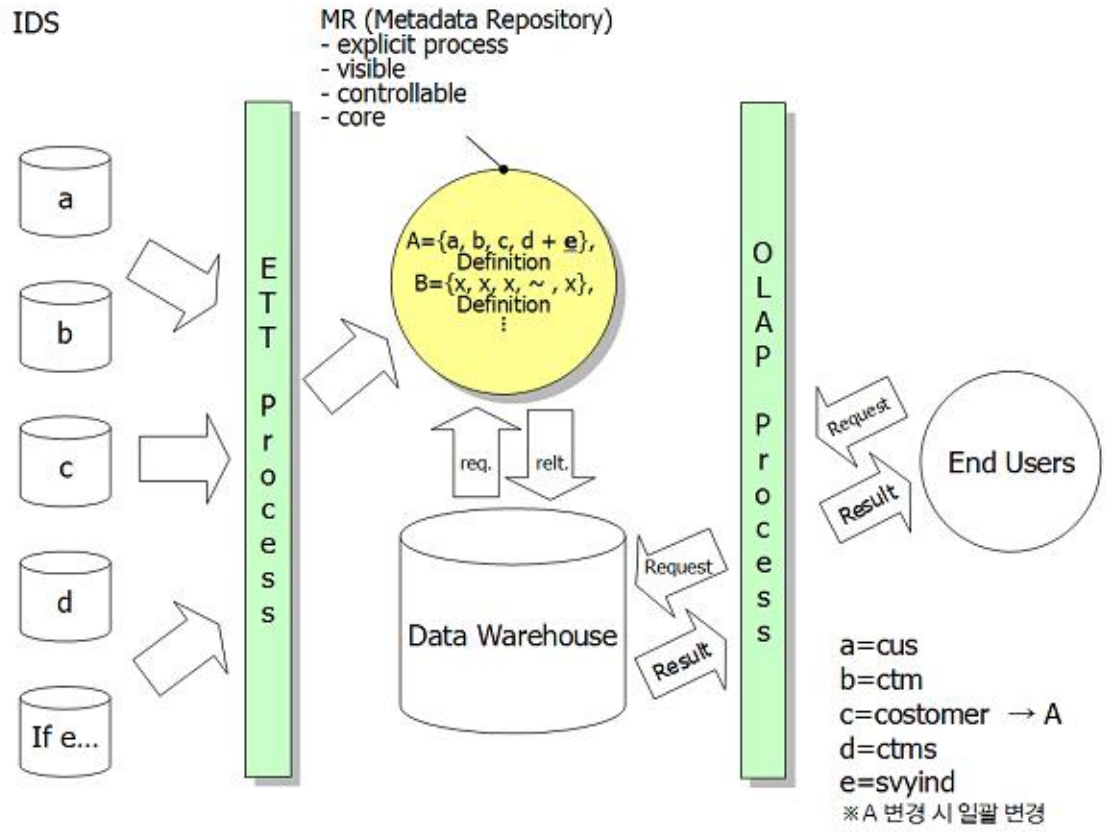
선행 연구 사례들을 종합한 메타데이터 관리 도구의 사용 현황은 다음과 같은 성격을 가진다.

1. 내재적 프로세스
2. 비시각적
3. 관리의 난점
4. 데이터의 참조 자료로서의 역할

3.2.2 통합 아키텍처

통합 아키텍처는 기존의 아키텍처와 다음과 같은 차이를 가진다(<그림 5>참조).

<그림 5> 메타데이터를 기반으로 한 데이터 통합 아키텍처



<그림 5>에서 메타데이터 관리 도구는 부서 간의 데이터들의 의미를 파악하고 있으며 이로 인해 통합된 전사적인 데이터 자원의 관리가 가능하다. 또한 메타데이터 관리 도구와 데이터 웨어하우스 간의 분리 구조로 인해 메타데이터의 관리가 용이하며 관리 도구를 통한 메타데이터의 일괄 수정 역시 가능하다. 신규 모델에 따르면 관리도구 내의 메타데이터의 변경이 정보 원천에서의 데이터 변경에 미치는 영향도 분석이 가능한데 이러한 분석을 통해 메타데이터의 관리에 있어 중요도에 따른 우선 순위를 파악할 수 있어 관리의 효과성 및 효율성을 추구할 수 있다.

통합 아키텍처의 메타데이터는 다음과 같은 특징을 가진다.

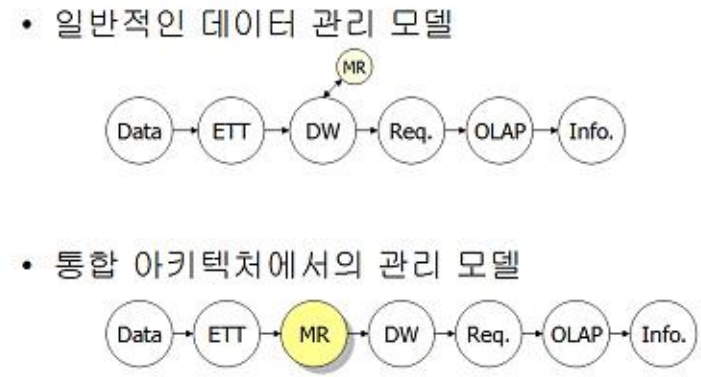
1. 외재적 프로세스
2. 가시성

3. 관리의 용이성

4. 데이터 관리에 있어 핵심(core) 역할

<그림 6>에서는 일반적인 데이터 관리 모델과 통합 아키텍처의 관리 모델에서의 데이터 처리 프로세스를 선형(linear)으로 비교하였다.

<그림 6> 데이터 관리 모델의 비교



즉, <그림 6>과 같이 일반적인 관리 모델의 경우 메타데이터는 참조되거나 참조되지 않을 수 있으나 통합 아키텍처에서는 메타데이터를 거치지 않고서는 데이터가 정보화 될 수 없다. 기존 모델에서는 정보 원천에서의 정보 발생에 대해 메타데이터와 관련된 제약을 주지 않는다. 각 부서들은 한시적으로 필요시 테이블을 생성하고 필드의 속성을 추가하며 이를 내부적으로 사용할 수 있다. 신규 모델에서는 이러한 부서간의 데이터 운용 행위가 메타데이터와 연관이 있을 경우 이를 메타데이터 관리 도구에 반드시 추가시키고 승인을 얻어야 한다. 즉 정보의 생성 원천에서의 정보 생성과 유통은 이를 관리하는 메타데이터 관리 도구의 허락을 득하여야 가능한 것이다. 따라서 메타데이터를 통한 관리의 경우 관리 대상의 범주가 확대되는 번거로움이 있지만 전사적인 데이터에 대한 이해, 활용 및 수정이 가능하게 된다.

Ⅳ. 결 론

본 연구에서는 선행 연구를 통해 데이터 자원 관리의 문제점을 지적하였다. 현재의 데이터 자원은 이의 효율적인 관리 방법의 부재로 인하여 점점 증가되는 그것의 중요성에도 불구하고 품질을 신뢰할 수 없는 경향이 있다. 이를 해결하기 위해 본 논문에서는 메타데이터

를 기반으로 한 데이터 통합 아키텍처를 제시하였다. 향후 신규 아키텍처를 통한 데이터베이스 관리의 경우 관리 범주가 증대되는데 이를 효율적으로 자동화 할 수 있는 방법에 대한 연구가 필요하다.

참고문헌

[1] 김정욱, 김영걸, "정보자원관리 관점에서의 통합 통제 아키텍처 메타데이터모형", 한국경영학회, 경영정보연구, Vol. 27 No. 4, pp. 875-890, 1998.

[2] 심부성, 고구진, 주종철, 박동인, 이필규, "지능형 메타데이터 시스템 설계 및 구현", 한국통신학회, 한국통신학회 학술대회 논문집, pp. 556-565, 1997.

[3] 이진수, 노희영, "B2B 전자상거래에서 메타데이터를 이용한 데이터베이스의 통합", 강원대학교 기초과학연구소, 기초과학연구 제13집, pp. 45 ~ 56, 2002.

[4] 조남철, 손명호, 김태훈, 이희석, "데이터웨어하우스 메타데이터 구축사례", 한국지능정보시스템학회, 학술대회발표, Vol. 1 No. 1, pp. 383-392, 1999.

[5] 한국데이터베이스진흥센터, "메타데이터 명세 및 메타모델의 표준화", 1998.

[6] Arun Sen, "Metadata Management: Past, Present and Future," Decision Support Systems, Vol 37, Issue 1, pp.151-173, 2004.

[7] Surajit Chaudhuri & Umeshwar Dayal, "An Overview of Data Warehousing and OLAP Technology," ACM Sigmod Record, Vol. 26, No. 1, pp.517-526, 1997.

[8] Vikram Ramchandra and Sreedhar Srikant, "Data Quality for Enterprise Risk Management," Business Intelligence Journal, Vol. 11, No. 2, 2006.

[9] Wayne W. Eckerson, "Mastering Metadata," Business Intelligence Journal, Vol. 9, No. 4, 2004.