제33회 한국정보처리학회 춘계학술발표대회 논문집 제17권 제1호 (2010. 4)

데이터웨어하우스를 위한 XMDR 기반의 데이터 정제시스템 설계

송홍율*, 첸드 아유시**, 정계동**, 최영근**
*광운대학교 유비쿼터스컴퓨팅학과
**광운대학교 컴퓨터소프트웨어학과
youry1029@chol.com

Design of data cleansing system based on XMDR for Datawarehouse

Hong-Youl Song*, Tsend Ayush**, Kye-Dong Jung**, Young-Keun Chol**
*Dept of Ubiquitous-Computing, Kwang-woon University
**Dept of Computer-Software, Kwang-woon University

요 약

데이터웨어하우스는 기업의 정책을 결정하는데 사용하고 있다. 그러나, 새로운 시스템이 추가되면 데이터 통합 측면에서 시스템간의 여러 가지 이질적인 특성으로 인해 많은 비용과 시간이 필요로 하게 된다. 따라서, 이러한 이질적인 특성을 해결하기 위해 데이터 구조의 이질성 및 데이터 표현의 이질성은 XMDR(eXtended Master Data Registry)를 이용하여 추상화된 쿼리를 생성하고, XMDR에 맞게 쿼리를 분리함으로써 이질성을 해결한다. 특히 본 논문에서는 XMDR을 이용하여 분산 시스템 통합시 로컬시스템의 영향을 최소화하고, 데이터웨어하우스의 정보를 실시간으로 생성하기 위해 분산된 환경에서 데이터 통합을 위한 표준화된 정보를 제공한다. 또한, 기존 시스템의 변경 없이 데이터를 통합하여 비용과 시간을 절감하고, 실시간 데이터 추출 및 정제 작업을 통해 일관성있는 실시간 정보를 생성하여 정보의 품질을 향상시킬수 있도록 한다.

1. 서론

데이터웨어 하우스에서는 데이터 통합을 위해 운영계 시스템의 데이터를 추출하여 정보계 시스템에서 활용하기 위한 자료를 생성하기 위한 방법으로 ETT(Extraction, Transformation, transportation)를 사용한다. ETT는 데이터를 추출하고 추출한 자료를 정제하여 정보계 시스템의 Fact Table에 자료를 생성하는 역할을 한다. 데이터웨어하우스는 과거로부터 현재까지 누적된 많은 데이터를 활용하기 때문에 시스템이나 환경에 맞게 새로운 데이터를 반복적, 주기적으로 데이터웨어하우스에 누적시켜야 한다. 생성주기가 길면 그 기간만큼 정보의 신뢰도는 떨어지게 되고, 짧으면 신뢰도는 높아지지만 시스템에 부하를 주게 된다. 오늘날 기업의 성장이 지속되고 글로벌화 되면서 IT시스템 또한 다양한 분야에서 여러 목적으로 사용됨에 따라, 시스템의 수가 증가하여 현재와 같은 복잡하고 다양한 IT 환경을 조성하게 되었다. 이런 복잡하고 다양한 IT 환경에서 최적의 프로세스를 효율적으로 지원하기 위해 전사의 동일한 기준으로 효율적인 기본 정보(Master Data)를 관리할 필요성이 높아지고 있다. 기준 정보는 일반적으로 회사 전 업무 영역과 연관된 프로세스에 사용되거나, 전사적인 프로세스를 통제하는데 사용되므로, 비즈니스 프로세스에 미치는 영향이 클 수 밖에 없다. 일반적인 MDM(Master Data Management) 시스템 기반의 데이터웨어하우스는 새로운 시스템의 추가시 데이터 이질성 및 기본 정보(Master Data) 통합의 문제들이 발생하여 확장이 어렵다. 이러한 문제를 해결하기위해 본 논문에서는 MDM 시스템의 확장된 기능으로 XMDR을 이용하여 데이터의 이질성을 해결하고, 일관된 기본 정보을 관리하여 각 로컬시스템으로부터 실시간으로 데이터를 추출 가공하여 데이터웨어 하우스에서 최신 정보를 사용할 수 있는 방안을 제시한다.

본 논문의 구성은 각 장별로 다음과 같다. 2장은 관련연구로 본 논문의 기반이 되는 ETT에 대하여 기술하고, 정제시스템의 근간이 되는 XMDR 구조에 대해 알아본다.

3장은 본 논문의 정제시스템에 대한 설계 방안을 기술하였고 적용예 및 비교분석을 수행한다. 4장에서는 결론 및 향후연구과제에 대해 기술하도록 한다.

2. 관련연구

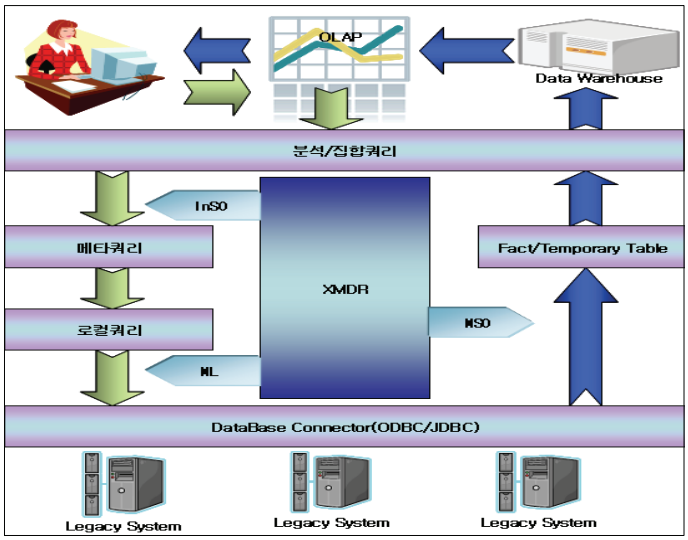
ETT(Extract Transfomation Transportation)는 데이터의 추출, 가공, 전송의 약자로 데이터를 소스시스템에서 추출하여 데이터웨어하우스에 로드시킨 상태에서 정제작업에 이르는 전 과정을 말하는 것으로 소스DB인 운영시스템 데이터를 변환, 정제하여 데이터웨어하우스에 적재하는 단계이다[1]. 이 작업을 위해서는 우선 데이터웨어하우스와 소스DB의 데이터를 대응(mapping)한 매핑표가 필요하다. 매핑표는 정보공유를 목적으로 데이터 표준화를 정립시킨 MDR(Meta Data Registry)로서, 소스DB의 데이터 구조와 변환, 정제 알고리즘에 대한 정보가 기록되어 있다[2].

XMDR은 MDR을 보다 더 확장한 개념으로 이전의 MDR에 전문 용어, 온톨로지, 시멘틱 웹(Semantic Web) 관리 기능을 확장한 것이다. XMDR은 데이터 통합에 따른 데이터 이질성을 해결하기 위하여 관계형 데이터베이스 메타데이터를 객체지향 데이터베이스에 저장하는 기술로 분산된 데이터의 이질성을 해결하고자 MDR(Metadata Registry)과 온톨로지를 결합하여 데이터를 통합하는 시스

템이다[3][4]. 또한, XMDR은 분산된 데이터 통합 또는 데이터웨어하우스에서 메타 데이터의 이질성을 해결하기 위한 메타 시멘틱 온톨로지와 메타 로케이션을 결합하고 실제 값들 사이의 이질성을 해결하기 위해 인스턴스 시멘틱 온톨로지를 결합한 것이다.
XMDR의 각 구성 요소들은 ISO/IEC 11179-3에서 제안한 데이터의 속성 명세를 따르며 메타 시멘틱 온톨로지(MSO:Meta Semantic Ontology)와 메타 로케이션(ML:Meta Location), 인스턴스 시멘틱 온톨로지(InSO:Instance Semantic Ontology)로 구성되어 있다[5].
본 논문에서는 XMDR의 구성요소를 데이터웨어하우스에 적용하여 데이터 추출 및 정제작업을 실시간 처리할 수 있도록 한다.

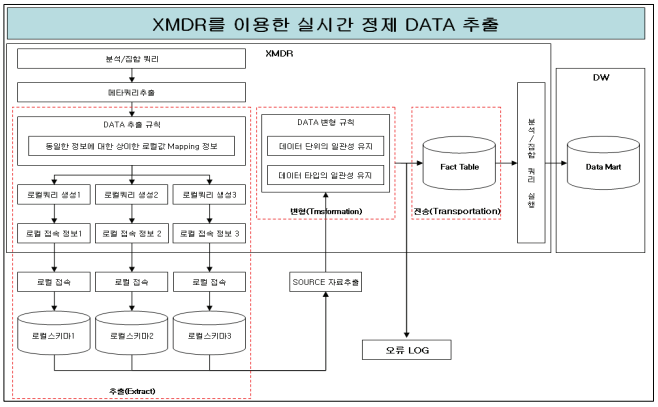
3. 데이터웨어하우스를 위한 XMDR 기반의 데이터 실시간 정제시스템

3.1 시스템 개요



(그림 1) 시스템 개요

일반적인 데이터웨어하우스에서는 사용자가 요청한 분석자료를 일괄 또는 배치형태의 스케줄 작업에 의해 생성된 데이터 마트로부터 자료를 조회하여 사용자에게 제공하므로 데이터마트에 자료를 적재한 후부터 현재 시점까지의 Gap이 발생할 수 있다. 발생한 Gap은 데이터의 품질에 영향을 주며 낮은 데이터 품질은 데이터의 신뢰성을 저하시켜 의사결정의 지연을 가져온다. XMDR기반의 데이터 정제시스템은 운영계 시스템별로 관리되는 데이터를 전사적 관점에서 통합관리할 수 있도록 하는 통합데이터베이스이다. 그림 1과 같이 XMDR 기반의 실시간 정제 시스템은 사용자가 요청한 자료의 질의를 분석하여 논리적인 메타쿼리로 변환한 후 메타쿼리로부터 분산된 로컬 시스템별로 로컬쿼리문을 변환한다. 변환된 로컬쿼리문은 XMDR의 ML정보를 이용하여 로컬시스템에 실시간 접속하여 질의를 실행하고, 추출한 자료는 정제 작업을 거쳐 자료를 통합함으로써 데이터 통합시 시간차이에서 오는 데이터의 Gap 발생을 최소화함으로써 사용자에게 최신의 분석 정보를 활용할 수 있도록 한다.
데이터웨어하우스에서 사용자가 요청하는 분석자료는 의사결정에 영향을 주는 시간과 공간 개념을 포함한 요약 데이터인 경우가 대부분이며 이런 차원테이블(Dimention Table)을 생성하기 위해서 각 로컬시스템에서 추출한 자료를 통합한 사실테이블(Fact Table)을 먼저 생성하여야 한다[6]. 이 시스템에서는 사용자에게 실시간 분석자료를 제공하기 위해 요청자료에 대한 질의문에 포함되어 있는 시간과 공간의 조건을 분석하여 각 로컬 시스템에서 데이터 추출 시 사용자가 요구하는 조건에 맞는 데이터들만을 추출할 수 있도록 한다, 사용자가 요구하는 데이터만을추출하므로 데이터 추출 속도 및 정제 작업의 시간을 단축하여 Fact Table에 자료를 생성한다.
예를 들어 2010년부터 현재시간까지 서울지역의 월별 판매금액을 조회하는 경우라면 '2010년부터 현재까지'라는 시간개념과 '서울'이라는 공간개념을 분석하여 로컬 시스템 자료 조회시 질의조건에 적합한 데이터들만 조회하여 Fact Table에 자료를 생성한 후 Fact Table로부터 사용자가 요청한 결과를 제공한다.



(그림 2) 시스템의 절차적 표현

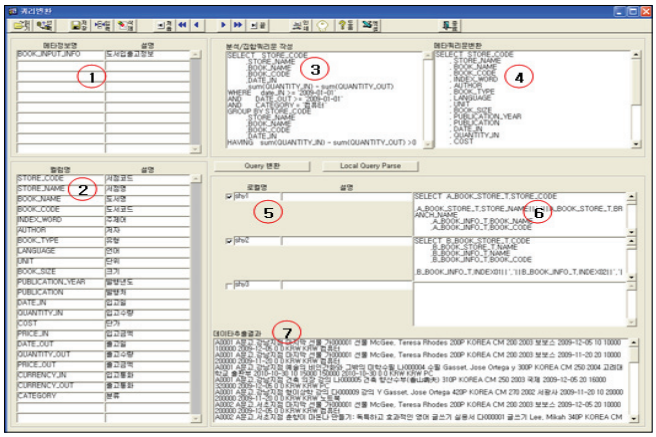
그림 2는 XMDR를 이용한 실시간 정제 DATA추출작업를 표현한 것으로 아래의 순서와 같이 진행된다.
Step 1. 사용자가 요구하는 Data Mart의 데이터를 추출할 분석/집합 쿼리를 분석한다.
Step 2. 분석/집합 쿼리로부터 메타쿼리를 추출한다. 메타쿼리는 Fact Table의 컬럼과 분석/집합쿼리의 Where절로 구성되어 있고, Where 절 구성시 XMDR의 InSO를 이용하여 분석/집합쿼리의 Where절을 확장한 형태이다.
Step 3. 메타쿼리를 컬럼 매핑 정보를 이용하여 로컬쿼리로 변환한다.
Step 4. XMDR의 ML 정보를 이용하여 로컬시스템에 접속한다.
Step 5.로컬시스템에서 로컬쿼리를 실행하여 데이터를 추출한다.
Step 6. 추출된 데이터를 XMDR의 MSO 정보를 이용하여 데이터를 정제한다.
Step 7. 추출 정제된 데이터를 Fact Table에 생성한다.
Step 8. Fact Table로부터 분석/집합 쿼리를 실행하여 Data Mart에 정보를 생성한다.

실시간 정제 시스템에서는 XMDR의 정보를 이용하여 분석/집합쿼리를 From절이 없는 메타쿼리로 변환하고 변환된 메타쿼리문을 메타컬럼과 로컬컬럼의 매핑맵을 이용하여 로컬 쿼리문으로로 변환한 후 각 로컬에 접속하여 자료를 Record 단위로 추출한다. 추출한 자료의 정제와 변형 작업을 거쳐 Fact Table에 자료를 생성하고, Fact Table로부터 분석/집합쿼리를 실행하여 정보를 제공한다.

3.2 적용예 및 비교분석

그림3은 분석집합쿼리를 이용하여 로컬시스템으로부터 자료를 추출한 예시화면으로 ①은 메타정보로서 추출 및 정제하고자 하는 자료들의 설명이다. ②는 메타정보의 논리적 컬럼으로 로컬 시스템의 물리적 스키마와 매핑되는 컬

럼들이다. ③은 예시 프로그램에서 작성한 분석/집합 쿼리이다. ④는 작성한 분석/집합쿼리를 쿼리변환기를 통해 변환된 From절이 없는 메타쿼리문의 결과이다. ⑤는 통합대상 로컬스키마 목록으로 XMDR의 ML정보를 포함하고 있다. ⑥은 메타쿼리를 로컬스키마별 로컬쿼리로 변환한 결과이다. ⑦은 ⑤에서 선택한 로컬 스키마별로 로컬쿼리문을 실행하여 데이터를 추출한 결과이다.



(그림 3) 쿼리변환 및 데이터추출 예시화면

표 1은 Oracle ODS와 본시스템을 비교분석한 결과이다.

<표 1> 관련시스템과의 비교분석

비교항목	Oracle ODS	본 시스템
데이터 통합방식	Data HUB를 이용하여 데이터 통합	XMDR을 통한 레거시 시스템을 직접 Acess하여 통합
확장성	수집정보 일원화 체계로 확장이 용이	XMDR을 통한 확장이 용이함
이질성극복	통합을 고려한 데이터 표준화 및 일관성 체계 확립	XMDR을 이용하여 데이터의 구문, 구조, 의미의 이질성 극복
일관성	레거시 시스템에서 발생하는 데이터의 변경은 Data HUB를 변경하여 일관성을 유지	레거시 시스템에서 발생하는 데이터의 변경은 XMDR Setup에 의해 일관성을 유지
검색효율성	데이터 허브 시스템을 이용한 검색 질의 수행	XMDR을 바탕으로 의미적으로 연관된 데이터를 포함한 검색 질의 수행

데이터 통합방식에서 Oracle ODS는 Data HUB를 이용하여 데이터를 중앙에 집중하는 방식으로 운영계 시스템이 증가할수록 데이터 허브에 대한 스키마와 추가된 운영계 시스템과의 연계를 새로이 구축해야 하는 단점이 있다. 본 시스템은 XMDR의 ML정보를 통하여 운영계 시스템에 직접 접속, 데이터를 추출하는 방식을 사용하였으며, 스키마의 이질성은 논리적인 메타데이터 스키마와 물리적인 운영계 시스템 스키마를 매핑하는 방식을 사용하여 새로운 운영계 시스템의 확장시 운영계 시스템의 변경을 최소화하여 유연하게 통합할 수 있도록 하였다. 데이터의 이질성에 대해서 Oracle ODS는 데이터 허브 시스템에서 지원하는 메타데이터를 이용하여 데이터의 구문적, 구조적 이질성을 극복하였고, 본 시스템에서는 XMDR의 MSO를 통하여 이질성을 극복하였다. 운영계 시스템의 변경에서 발생하는 데이터의 일관성은 Oracle ODS는 Data HUB의 변경을 통해 일관성을 유지하였고, 본 시스템에서는 데이터 통합방식에서 제시한 것과 같이 XMDR의 논리적 메타데이타 스키마와 변경된 운영계 시스템의 물리적 스키마와의 매핑으로 일관성을 유지할 수 있도록 하였다.

4. 결론

본 논문에서 설계 및 구현된 시스템은 분산되어 있는 데이터베이스 시스템으로부터 실시간 데이터 추출 및 정제 작업을 통해 통합 데이터를 제공하기 위하여 표준 스키마 문서와 XMDR, 쿼리분석 및 변환, 메타데이터 사전을 이용하여 통합 시 발생할 수 있는 이질적인 문제를 해결하였고 사용자의 요청에 실시간으로 데이터를 통합할 수 있게 구현한 시스템이다. XMDR기반의 실시간 데이터 정제 시스템의 장점은 분산된 환경에서 데이터를 통합하기 위한 표준화된 정보를 제공, 일관성있는 정보를 생성할수 있도록하여 정보의 품질을 향상시킨다. 또한 새로운 분산환경의 통합시 로컬 시스템의 변경을 최소화하고, XMDR에 로컬 정보를 추가함으로써 데이터웨어하우스를 위한 정보를 통합할 수 있도록 해 준다.

실시간 정제시스템의 응답시간은 데이터의 양에 비례한다. 제안된 시스템에서는 실시간처리를 위해 Data Mart에 데이터를 생성하기 위한 쿼리를 분석하여 추출하는 데이터의 범위를 최소화 하도록 하였다. 그러나 실제 Data Mart의 기본이 되는 Fact Table을 생성하기 위한 추출 데이터가 대용량이라면 실시간 정제시스템의 응답시간은 그 만큼 늦어질 수 밖에 없다. 따라서 앞으로 대용량 기반의 실시간 데이터 추출 및 정제 시간을 단축하기 위한 방법에 관한 연구가 필요하다.

참고문헌

[1] 여성주, 왕지남, “데이터웨어하우스에서 이질적 형태를 가진 데이터의 추출을 위한 Extraction”, 산업경영시스템학회지, 2001

[2] Andrew White, David Newman, Debra Logan, John Radcliffe, “Mastering Master Data Management”, Garther, 2006

[3] 정계동, 황치곤, 최영근, “분산 환경에서 XMDR을 이용한 예약 정보 시스템”, 한국해양정보통신학회논문지 Vol.11 No.10 pp.1957-1967, 2007

[4] 국윤규, “하이브리드 에이전트에 의한 효율적인 데이터 그리드 시스템”, 광운대학교 박사학위 논문, 2006.

[5] Kevin D. Keck and John L. McCarthy, “XMDR: Proposed Prototype Architecture Version 1.01”, http://www.xmdr.org, February 3, 2005

[6] 박시우, 지원철, “사례기반추론을 이용한 데이터 웨어하우스 설계방법론”, 대한산업공학회/한국경영과학회 ‘98 춘계공동학술대회 논문집, 1998