http://dx.doi.org/10.14400/JDC.2014.12.2.263

클라우드 환경에서 XMDR-DAI를 이용한 데이터 정제 시스템

문석재*, 정계동*, 이종용*, 최영근*
광운대학교*

Data Cleaning System using XMDR-DAI in Cloud

Seok-Jae Moon*, Kye-Dong Jeong*, Jong-Yong Lee*, Young-Keun Choi*
Department of Computer Science Kwangwoon University*

요 약 클라우드 환경에서 비즈니스 인텔리전스를 위한 DW(Data Warehouse)는 기업 내에 데이터를 의사결정, 기업 정책을 결정하는데 사용하고 있다. 그러나 클라우드 환경에서 새로운 시스템이 추가되면 데이터 통합 측면에서 시스템간의 여러 가지 이질적인 특성으로 인해 많은 비용과 시간이 필요로 하게 된다. 따라서 본 논문에서는 클라우드 환경에서 비즈니스 인텔리전스를 위한 데이터 정제 시스템을 제안한다. 제안 시스템은 XMDR-DAI를 이용하여 분산된 시스템을 통합할 때 로컬 시스템의 영향을 최소화하고, DW의 정보를 실시간으로 생성하기 위해 데이터 통합을 위한 표준화된 정보를 제공한다. 또한 기존 시스템의 변경 없이 데이터를 통합하여 비용과 시간을 절감하고, 실시간 데이터 추출 및 정제 작업을 통한 일관성 있는 실시간 정보를 생성하여 정보의 품질의 향상시킬 수 있도록 한다.

주제어 : 클라우드 컴퓨팅, 메타데이터, 비즈니스 인텔리전스, 데이터웨어하우스, XMDR, 데이터 정제

Abstract In cloud environment, business intelligence data warehouse is used for decision making and enterprise policy. But if new system is added in cloud environment, much cost and time is needed due to heterogenous characteristics in data integration. This paper suggests a data cleaning system for business intelligence in cloud environment. The proposed system minimizes the effect of local system when it integrates distributed system using XMDR-DAI. And this system provides standardized information to generate information of data warehouse in real time. Also the proposed system saves cost and time by integrating the data without a change of existed system. And it can improve quality of information by generating coherent information through data extraction and cleaning work in real time.

Key Words : Cloud Computing, MetaData, Business Intelligence, Data Warehouse, XMDR, Data Cleaning

1. 서론

클라우드 환경[1]에서 비즈니스 인텔리전스[2]는 다른 아키텍처를 가진 레거시 시스템 간 상호 작동의 필요성이 점점 중요해지는 기업에서 지속적으로 확대되고 있다. 이에 기업들은 비즈니스 인텔리전스를 이용하여 의사결정을 지원하기 위해 데이터를 통합하는 솔루션으로는 DW(Data Warehouse)를 사용하고 있다. DW는 로컬 시

* 본 논문은 2013년 광운대학교 연구비에 의하여 지원되었음
Received 30 December 2013, Revised 7 February 2014
Accepted 20 February 2014
Corresponding Author: Seok-Jae Moon(Department of Computer Science)
Email: msj8086@kw.ac.kr

ISSN: 1738-1916

스템에서 발생하여 누적된 데이터로부터 기업이 필요한 정보를 추출하는데 효과적이기 때문이다. 그에 따른 기업들의 성장이 지속되었지만, 글로벌화 되면서 로컬 시스템들이 수가 증가하고 복잡해졌다. 이런 환경에서 기업은 최적의 프로세스를 효율적으로 지원하기 위해 전사의 동일한 기준으로 효율적인 기준 정보(Master Data)를 관리할 필요성이 높아지고 있다[3]. 기준 정보는 일반적으로 회사 전 업무 영역과 연관된 프로세스에 미치는 영향이 클 수밖에 없다. 정확한 기준 정보에 의해 회사 업무가 진행되면, 동일한 기준으로 프로세스를 처리할 수 있고, 비효율적인 반복 작업을 제거하여 최적화된 프로세스를 통해 트랜잭션을 처리할 수 있다[3]. 그러나 일반적인 MD시스템 기반의 DW는 새로운 시스템 추가 시 데이터의 이질성 문제, 기준 정보의 문제들이 발생하여 확장이 어렵다. 본 논문에서는 클라우드 환경에서 비즈니스 인텔리전스를 위한 XMDR-DAI 기반의 데이터 클렌징 시스템을 제안한다. 제안 시스템은 마스터 정보 시스템의 확장된 기능으로 XMDR[4]을 이용하여, 데이터의 이질성을 해결한다. 그리고 일관된 기준 정보 관리로 각 로컬 시스템으로부터 실시간으로 데이터를 추출 가공하여 DW에서 최신 정보를 사용할 수 있는 방안을 제시한다. 본 시스템의 목적은 XMDR-DAI[5]를 이용하여 비즈니스 인텔리전스를 위한 시스템 데이터 통합 시 로컬 시스템의 영향을 최소화하고, DW의 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 통한 일관성 있는 실시간 정보를 생성하여 정보의 품질을 향상시킬 수 있도록 한다. 본 논문의 구성은 2장은 관련연구를 기술하고, 3장은 제안 시스템에 대해서 기술한다. 4장에서는 시스템 적용을 5장에서는 결론에 대해 기술한다.

2. 관련연구

2.1 클라우드 환경

클라우드 컴퓨팅은 유틸리티 컴퓨팅처럼 기업이나 사용자가 실제로 소유하지 않은 소프트웨어나 플랫폼, 심지어 하부 구조까지도 필요한 만큼 동적으로 빌려 쓰고, 사용한 만큼 비용을 지불하는 방식을 사용한다. 따라서 앞서 언급한 작업 부하의 변동 폭이 큰 응용 프로그램의 서비스 질을 개선하는 좋은 방법이 될 수 있다. 클라우드 컴퓨팅은 새로운 패러타임이라기보다 과거의 그리드 컴퓨팅, 유틸리티 컴퓨팅, 병렬 처리 및 분산 시스템의 등의 특징이 복합되어 있는 패러타임이라 할 수 있다[6]. 클라우드 컴퓨팅 환경은 다양한 구조를 가질 수 있지만 기본적인 구성요소인 데이터 센터의 구조는 클라이언트, 웹 서버와 응용 서버, 그리고 데이터베이스 서버를 포함하는 서버층, 그리고 저장장치로 구성된다. 데이터 센터 안에는 여러 개의 클러스터가 존재하고, 각 클러스터에는 다수의 서버로 구성된 랙이 있다. 각 데이터 센터들은 인터넷 망을 연결됨으로써 클라우드 컴퓨팅 환경을 구성하게 된다. 클라우드 컴퓨팅 환경이 과거의 컴퓨팅 환경과 구별되는 특징들은 분산/병렬처리, 확장성, 유연성, 높은 가용성과 낮은 지연, 다양한 수준의 QoS 등이다. 최근 가상화 등 클라우드의 핵심 기술들에 대한 연구가 활발한 가운데, 클라우드가 활성화되기 위해 선결되어야 할 문제점과 고려사항에 대한 연구도 활발하다. 구체적인 문제점들로는 서비스 가용성과 데이터 기밀성, 전송 지연과 같은 기술적인 측면에서의 고려 사항과 안정성, 데이터 보안성, 정보 유출에 대한 이용자 측면에서의 고려사항 등이 있다.

2.2 XMDR-DAI

본 XMDR-DAI[5]의 정의 및 구성요소는 다음과 같다.

- Meta-Semantic Ontology(MSO): MSO는 로컬 데이터베이스들의 메타데이터 스키마 정보를 시소러스화한 것으로, 표준으로 지정한 메타데이터 스키마에 매핑하여 메타데이터간의 관계성과 이질적인 충돌 문제를 해결할 목적으로 정의한 것이다. 또한 MSO는 스키마 표준인 글로벌 메타데이터 스키마를 로컬 메타데이터 스키마로 변환하기 위해서 매핑할 때 필요한 정보이다.
- Meta-Location(MLoc): MLoc는 MSO와 연계하여 로컬 데이터베이스들의 물리적인 위치, 접근권한 정보 등을 등록한 것이다. 이는 데이터 접근 및 통합에서의 상호운용상에 필요한 데이터 이주 및 트랜잭션 과정을 비즈니스 프로세스 메시지가 해당 위치에 전달될 수 있도록 정의한 것이다.
- Instance-Semantic Ontology(InSO): InSO는 실제

인스턴스 값(value)간의 연관성(association) 정보를 매핑 구성하여 시소러스화 한 것으로, 인스턴스 값 사이에 의미성, 유사성, 유효성을 고려하여 정의한 것이다. 예를 들어, 단위, 형식(ex: mm->cm, kg->pound, mm/dd/yy->yy-mm-dd)의 불일치를 충돌 정보를 분류하여 정의한 것이다.

- MetData Registry(MDR): MDR은 각 로컬 데이터베이스의 메타데이터 개체 스키마를 등록하여 관리하는 것이다. 이 MDR은 글로벌 스키마 영역과 로컬 스키마 영역으로 구성된다.
  - Global Schema: 글로벌 스키마는 각 로컬 스키마들의 표준 스키마를 선정하고, 이를 비즈니스 협업에 맞게 구성한 것이다.
  - Local Schema: 로컬 스키마는 협업에 참여하는 로컬 데이터베이스의 스키마를 등록한 것이다.

위와 같이 정의된 XMDR[4] 요소들은 ISO/IEC 11179-3에 기술된 데이터 속성 명세를 따른 것이다. 데이터의 기본 속성은 식별속성, 정의속성, 관계속성, 표현속성으로 명세는 다음과 같이 정의하였다.

- Identify attribute: 데이터 요소의 식별을 위한 속성.
- Define attribute: 데이터 요소의 의미를 갖는 속성.
- Presentation attribute: 데이터 요소의 표현 방식을 위한 속성.

2.3 ETT

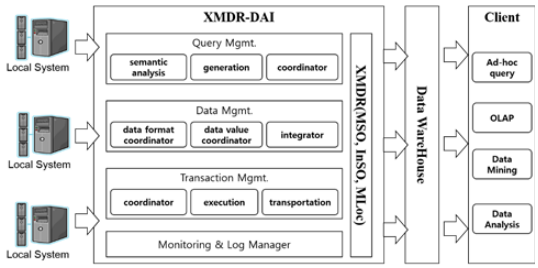
ETT(Extract Transfomation Transportation)[7]는 데이터를 소스시스템에서 추출하여 DW에 적재 시킨 상태에서 정제작업까지 이르는 전 과정을 말하는 것으로 소스DB인 운영시스템 데이터를 변환, 정제하여 DW에 적재하는 단계들이다. 이 작업을 위해서는 우선 DW와 소스DB의 데이터를 매핑한 매핑표가 필요하다. 매핑표에는 소스DB의 데이터 구조와 변환, 정제 알고리즘에 대한 정보가 기록된다. 추출에는 정보계 데이터의 성격에 따라 로컬의 데이터를 DW에 초기 적재하는 이행(Population) 작업과 로컬에서 생성, 갱신, 삭제된 정보가 운영 시점에서 DW에 반영되는 전송(Transmission) 작업의 두 가지 유형이 있다. DW에서 필요한 최종 테이블은 실적데이터를 가지고 있는 사실테이블과 공통적인 데이터를 관리하는 차원 테이블이 있다. ETT의 형태는 사실 테이블과 차원 테이블을 어떻게 설계하여 만들어 주느냐에 따라 시스템의 부하가 결정된다. 기간계 시스템에서 대부분의 차원테이블을 작성하여 DW쪽으로 전송한다면 DW의 역할은 그 만큼 간단해 지고 부하 또한 감소된다. 하지만 로컬 시스템의 부하를 무시할 수 없으므로 시스템 구현 시 로컬의 부하를 고려, trade-off관계를 잘 규명하여 결정하도록 하여야 한다. 일반적으로 로컬 시스템의 경우 실제 업무가 운영되는 곳이므로 시스템의 부하로 인한 업무의 지연을 잘 고려하여야 한다.

3. 제안 시스템

3.1 시스템 개요

일반적인 DW에서는 사용자가 요청한 분석 자료를 일괄 또는 배치형태의 스케줄 작업에 의해 생성된 데이터 마트로부터 자료를 조회한다. 그 조회한 결과를 사용자에게 제공하므로 데이터 마트에 자료를 적재한 후부터 현재 시점까지의 Gap이 발생할 수 있다. 발생한 Gap은 데이터의 품질에 영향을 주며, 낮은 데이터 품질은 신뢰성을 저하시켜 의사결정의 지연을 가져온다.

본 XMDR-DAI 기반의 데이터 정제 시스템은 사용자가 요청한 자료 추출 쿼리문을 분석하여 클라우드 환경에서 분산된 여러 로컬 시스템에서 실행할 수 있는 쿼리으로 변환한다. 변환된 쿼리문은 각 로컬 시스템에 실시간 접속하여 실행한다. 실행 후, 추출된 자료들은 정제 작업을 통해 표준화된 형태로 통합하여 데이터 발생 시 시간 차이에서 오는 데이터의 Gap 발생을 최소화함으로써 사용자에게 최신의 분석 정보를 활용할 수 있도록 한다. 본 시스템은 각 로컬 시스템 에서 운용되는 데이터를 전사적 관점에서 통합할 수 있도록 하는 통합형 데이터베이스이다. DW에서 사용자가 요청하는 분석자료는 의사결정에 영향을 주는 시간과 공간 개념을 포함한 요약 데이터인 경우가 대부분이며, 차원테이블(Dimention Table)을 생성하기 위해서는 각 로컬에서 추출한 자료를 통합한 Fact 테이블을 먼저 생성하여야 한다. 본 시스템에서는 실시간 데이터 추출 및 정제 작업을 위해 XMDR-DAI를 구성하였으며, XMDR-DAI는 Query Mgmt.(QM), Data Mgmt.(DM), Transaction Mgmt.(TM), Monitoring & Log Manager로 [Fig. 1]과 같이 Agent 기반으로 구성된다.



[Fig. 1] Propose System

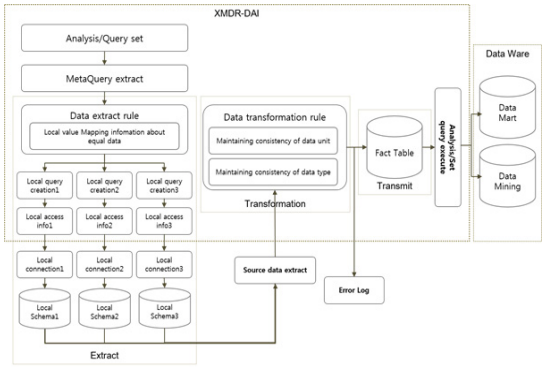
• Query Mgmt.(QM): QM은 사용자가 요청한 질의를 분석하여 논리적인 메타쿼리로 변환된다. 변환하는 경우 의미 이질성을 가진 조건들은 조정자(Coodinator)에 의해 확장된 조건으로 변환된다. 이 메타쿼리는 XMDR의 매핑맵을 통해 로컬쿼리로 다시 변환된다.

• Transaction Mgmt.(TM): TM은 로컬 시스템으로부터 추출한 자료를 DM에게 제공한다. DM에 의해 정제 및 통합된 데이터는 DW로 전송한다.

• Data Mgmt.(DM): DM은 각 로컬 시스템으로부터 추출한 자료를 통합하고, 통합된 자료들 중 의미 이질성, 표현 및 정의 이질성을 가진 데이터는 Data format coordinator, Data value corrdinator에 의해 일원화된 데이터로 조정된다.

• Monitor & Log Manager(MLM): MLM은 각 에이전트의 활동을 감시하며, 예외발생 시 로그를 생성하여, 관리자에게 통보하는 역할을 수행한다.

XMDR-DAI는 사용자에게 실시간 분석자료를 제공하기 위해 요청자료에 대한 쿼리문에 포함되어 있는 시간과 공간의 조건을 분석하여 각 로컬 시스템에서 데이터 추출 시 해당 조건에 적합한 데이터들만을 추출하여 Fact Table에 생성하도록 한다. 예를 들어 '2013년부터 현재시간까지 서울지역의 월별 판매금액을 조회'하는 경우라면 '2013년부터 현재까지'라는 시간개념과 '서울'이라는 공간개념을 분석하여 질의 조건을 도출한다. 도출된 질의 조건을 활용하여 로컬 시스템 쿼리문을 생성한다. 질의 조건에 적합하게 추출된 데이터들을 Fact Table에 생성하고 Fact table로부터 사용자가 요청한 결과를 제공한다. XMDR-DAI를 이용한 실시간 정제 데이터 추출작업은 [Fig. 2]과 같이 진행된다.



[Fig. 2] System Sequence

Step 1. 사용자가 요구하는 Data Mart의 데이터를 추출할 분석/집합 쿼리를 분석한다.

Step 2. 분석/집합 쿼리로부터 메타쿼리를 추출한다. 메타쿼리는 메타정보의 논리적 컬럼과 분석/집합쿼리의 조건절로 구성되어 있고, 메타쿼리의 조건절은 XMDR의 InSO를 이용하여 분석/집합쿼리의 조건절을 확장한 형태이다.

Step 3. 메타쿼리를 컬럼 매핑 정보를 이용하여 로컬 쿼리로 변환한다.

Step 4. XMDR-DAI의 ML 정보를 이용하여 로컬시스템에 접속한다.

Step 5. 로컬시스템에서 쿼리를 실행하여 데이터를 추출한다.

Step 6. 추출된 데이터를 XMDR-DAI의 MSO 정보를 이용하여 데이터를 정제한다.

Step 7. 추출 정제된 데이터를 Fact(Temp) Table에 생성한다.

Step 8. Fact(Temp) Table로부터 분석/집합 쿼리를 실행하여 Data Mart에 정보를 생성한다.

XMDR-DAI 서버는 사용자가 요청하는 쿼리문을 분석하여 질의 조건이 확장된 메타쿼리로 변환하고, 메타컬럼과 로컬 컬럼의 매핑맵을 이용 쿼리로 변환한 후, 각 로컬시스템에 접속하여 자료를 추출, 추출된 자료의 정제 및 변형 작업을 거쳐 사용자가 요청한 결과를 제공하는 기능을 수행한다.

3.2 데이터 추출

데이터 추출 속도는 일반적으로 데이터의 양과 비례

한다. [Fig. 1]의 QM은 전체 데이터가 아닌 사용자가 요청하는 데이터만을 추출하여 추출속도를 향상시키기 위해 구문 분석기를 통해 사용자가 요청한 분석/집합 쿼리를 형태로소 분리한 후에 메타정보, 예약어, 컬럼, 온톨로지 항목, 상수 등을 추출한다. 추출된 메타 정보로부터 쿼리변환기는 메타정보의 매핑 컬럼을 조회하여 From절이 없는 논리적인 메타쿼리로 변환한다. 메타쿼리로 변환 시 Coodinator은 XMDR-DAI의 InSO를 이용하여 조건절에서 추출한 정제 항목의 관계속성에 해당하는 메타자료조건 매핑맵으로부터 상수 값과 동일한 의미를 가진 상수 값들을 추출한다. 추출한 상수 값들을 조건절에 포함시켜 사용자가 요청한 자료 추출 시 의미적 이질성을 가진 전체 데이터를 조회할 수 있도록 하여 의미적 이질성을 극복하도록 한다. 관계속성은 객체지향설계 방법의 Is-a, Has-a, Equal으로 관계성을 분류하여 구성하였다. 다음 [Fig. 3]는 XMDR-DAI 내부의 InSO 저장 프로시저 알고리즘이다.

```
Function f_get_isno_value(p_value Varchar2) Return Varchar2 Is
Begin
//input value : select condition
retval := '(''||p_value||'''';
//input value and Equal-a relation value select
For rec_equal in c_equal(p_value) Loop;
   retval := retval||', ''||rec_equal.child_value||'''';
End Loop;
//input value and Has-a relation value select
For rec_has in c_has(p_value) Loop
   retval := retval||', ''||rec_has.child_value||'''';
   //Has-a releative and Equal relation value select
   For rec_equal in c_equal(rec_has.child_value) Loop;
      retval := retval||', ''||rec_equal.child_value||'''';
   End Loop;
   //selected data and Is-a relation value select
   For rec_is in c_is(rec_has.child_value) Loop
      retval := retval||', ''||rec_has.child_value||'''';
      //Is-a relation and Equal relation value select
      For rec_equal in c_equal(rec_is.child_value) Loop;
         retval := retval||', ''||rec_equal.child_value||'''';
      End Loop;
   End Loop;
End Loop;
retval := retval || ')';
Return retval;
End f_get_isno_value;
```

[Fig. 3] Stored procedure Algorithm, of InSO

3.3 데이터 정제 및 변형

3.3.1 스키마 정제

메타 쿼리의 구문형식은 "Select select-list Where conditions"의 형태를 지니고 있다. 메타쿼리에 From절을 명시하지 않는 이유는 구문분석기에서 추출한 메타정보와 매핑된 로컬시스템의 테이블정보를 로컬쿼리의 From절로 명시하기 때문이다. 논리적인 메타정보와 물리적인 로컬시스템 테이블과의 관계를 정의하여 매핑맵을 생성함으로써 새로운 로컬 시스템 추가 시 로컬시스템의 변경 없이 매핑 관계의 정의만으로도 확장이 가능하게 한다. 메타쿼리는 로컬시스템으로부터 최하위 레벨의 소스 데이터를 추출하기 위한 논리적인 쿼리로 분석 및 집합 구문을 사용하지 않는다. 스키마 정제는 논리적인 메타쿼리를 물리적 로컬 스키마에서 질의를 수행할 수 있도록 로컬 스키마에서 질의를 수행할 수 있도록 로컬쿼리를 생성하는 작업이다. 이를 위해서 스키마 정제 마스터 테이블의 항목을 메타정보, 메타컬럼 정보, 로컬정보, 로컬테이블 정보, 로컬컬럼 정보, 테이블매핑 정보, 컬럼매핑 정보 구분하여 등록하였다. 각 테이블에 로컬정보 및 로컬 테이블과 컬럼 정보를 등록하고, 전사적 관점에서 통합할 메타정보를 등록한 후 메타정보의 컬럼과 로컬스키마 컬럼을 매핑하여 유연하게 메타쿼리를 로컬쿼리로 변경할 수 있도록 한다. 메타쿼리는 구문분석기을 통해 Select절, Where절로 구분하고 각 구문별 구문분석을 통해 메타컬럼, 예약어, 상수, 기호 등을 분리한다. 분석된 메타컬럼은 컬럼매핑정보를 통해 로컬컬럼으로 변환하고, 각종 예약어는 각 로컬스키마에서 사용하는 예약어로 변환한다. 변환된 로컬스키마별 조건절에 테이블매핑정보에 있는 로컬테이블의 기본 조건절을 포함시켜 조건절을 생성한다. 변환된 Select절과 Where절을 연결하여 로컬쿼리를 생성한다.

3.3.2 데이터 정제

데이터 정제는 소스데이터의 형식, 내용을 검증하여 가치 있는 데이터로 만드는 과정이다. 데이터는 로컬스키마 별로 데이터타입, 길이, 날짜 형식, 수치에 대한 단위, 환율 등이 서로 상이할 수 있으므로 DW에서 사용하기 위해서는 일원화된 정보로 정제하여야 한다. 정제속성은 병합, 일자, 통화, 단위 4가지로 구분하였다. 병합은 메타컬럼이 로컬스키마에서 여러 개의 컬럼으로 있는 겨우 로컬 쿼리를 생성시 컬럼 연결자를 사용하여 하나의 컬럼으로 인식할 수 있도록 하여 데이터 구조에서 오는 이질성을 해결한다. 일자는 메타컬럼의 타입과 일자표현 형식이 로컬 시스템과 상이한 경우 로컬시스템의 일자컬럼을 메타컬럼과 동일한 표현 방식으로 변환하여 데이터 정의 및 표현에서 오는 이질성을 해결한다. 단위는 여러 단위로 표현된 단위의 가치를 표준 단위로 변환하여 데

이터의 의미 이질성을 해결한다. 다음 [Fig. 4]는 데이터 이질성 정제 알고리즘이다.

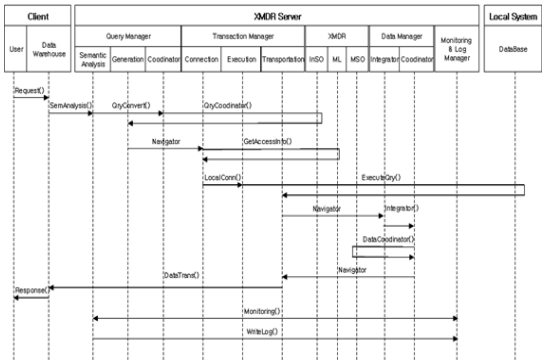
```
If meta_data_type<>local_data_type Then
 If meta_dat2_format<>local_dat2_Format THEN
  If meta_data_type='DATE' then
    select-column = to_date(local_column_name, local_data_format)
  Else If local_data_type='DATE' Then
      select-column=to_char(local_column_name
                  ,meta_data_format)
   Else local_data_type = 'CHAR' Then
      select-column = to_char(to_char( to_date(local_column_name
                          ,local_date_format),meta_date_format)
  End If;
 End If
End If;
```

[Fig. 4] Heterogeneity of data refinement algorithm

3.5 작업 흐름

[Fig. 5]는 사용자가 DW에 요청한 정보에 대하여 XMDR-DAI가 로컬시스템으로부터 자료를 추출하여,결과를 전송하는 작업흐름이다.

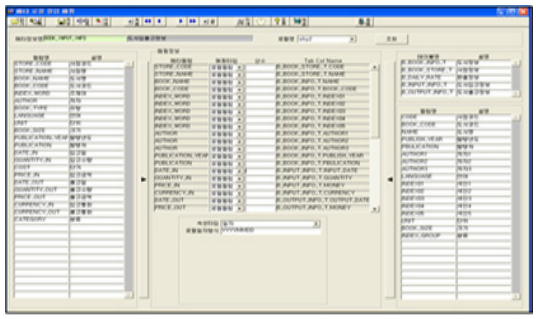
클라이언트, XMDR-DAI 그리고 로컬 시스템부분들을 분할하여 수행하는 과정을 보였다. 각 과정의 작업들은 다음과 같다.



[Fig. 5] Work Flow

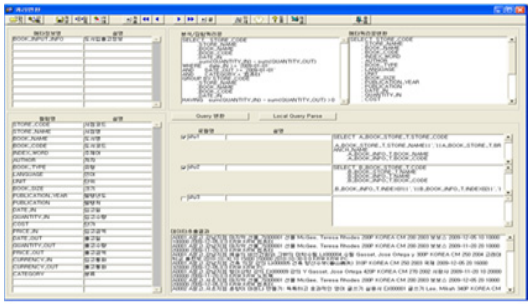
4. 시스템 적용

[Fig. 6]은 메타 정보 및 메타컬럼 정보를 등록하는 인터페이스이다. 인스턴스 매핑명은 XMDR-DAI의 InSO를 위한 정보로서 의미 이질성의 값을 가진 컬럼에 해당하는 인스턴스 매핑정보를 등록하여 메타쿼리 변환할 때 메타인스턴스 매핑 정보로부터 데이터를 조회하여 조건절을 확장하는데 사용한다. 메타컬럼 등록시 속성 타입은 통화와 단위 두 가지로 구분하였다. 통화는 데이터 정정 작업할 때 기준일자 컬럼명과 통화 컬럼명 변경할 통화를 입력 값으로 받아 표준통화금액을 계산하는데 사용한다. 단위는 대상단위가 등록되어 있는 컬럼명과 변경할 단위를 입력 값으로 받아 표준단위로 계산하는데 사용한다. 로컬 시스템으로부터 데이터를 추출하여 Fact Table에 생성한 후, 해당 컬럼의 속성이 통화 또는 단위일 경우 해당 정보를 이용하여 표준통화 및 표준단위로 데이터를 정제한다.



[Fig. 6] Local column mapping metadata

[Fig. 7]에서 메타컬럼 등록시의 속성타입과의 차이점은 메타컬럼 등록시 속성타입은 로컬에서 자료를 추출하여 Fact Table에 자료를 저장한 후 데이터를 변형하는 것이다.



[Fig. 7] Query translation and data extraction

5. 결론

본 논문에서 설계 및 구현된 시스템은 클라우드 환경에서 분산되어 있는 DBMS으로부터 XMDR-DAI의 표준 스키마 문서와 쿼리분석 및 변환, 메타데이터 사전을 이용하여 통합 시 발생할 수 있는 이질적인 문제를 해결하고 사용자의 요청 즉시 통합된 데이터를 제공할 수 있게 구현한 시스템이다. 통합을 위해 먼저 분산 환경에서 발생할 수 있는 이질적인 문제를 다음과 같이 정의하여 해결하였다. 첫 번째, 데이터 구조의 이질성은 표준 스키마 문서와 로컬 데이터베이스의 접근 및 구조 정보를 사전에 XMDR-DAI에 등록하고, 쿼리 작성 및 분석기를 이용하여 Data Mart를 생성하기 위한 쿼리는 표준 스키마 문서를 기준으로 하여 메타쿼리로 변환하였다. 두 번째, 메타 쿼리는 로컬 데이터베이스 시스템의 쿼리로 변환하여 구조적인 이질성으로 인한 데이터 수집에 따른 문제를 해결하였다. 또한 데이터 정의와 데이터 표현의 이질성은 표준 스키마 문서를 기준으로 작성한 유사어 사전인 메타데이터 사전을 이용하여 정의 및 표현의 이질성을 해결하였다. 본 논문에서 제안한 XMDR-DAI 기반의 실시간 데이터 정제 시스템의 장점은 분산된 환경에서 데이터를 통합하기 위한 표준화된 정보를 제공함으로써 일관성 있는 정보를 생성할 수 있도록 하고, 정보관리의 일원화를 통해 정보의 품질을 증대시킨다. 또한 새로운 분산환경의 통합 시 로컬 시스템의 변경 없이 XMDR-DAI에 정보를 추가함으로써 DW를 위한 데이터 통합을 할 수 있도록 해준다. 그리고, 데이터 통합을 위한 시스템의 변경작업을 최소화함으로써 비용과 시간이 절감되는 효과를 얻을 수 있다. 실시간 정제시스템의 응답시간은 데이터의 양에 비례한다. 제안된 시스템에서는 실시간 처리를 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 데이터 추출 및 정제 시간을 단축하기 위한 방법에 관한 연구가 필요하다.

ACKNOWLEDGMENTS

This present has been conducted by the research grant of Kwangwoon University in 2013.

REFERENCES

[1] Armbrust, Michael, et al. "A view of cloud computing." Communications of the ACM 53.4 (2010): 50-58.

[2] Liautaud, Bernard, and Mark Hammond. e-Business intelligence: turning information into knowledge into profit. McGraw-Hill, Inc., 2000.

[3] Smith, Brian L., and Simona Babiceanu. "Investigation of extraction, transformation, and loading techniques for traffic data warehouses." Transportation Research Record: Journal of the Transportation Research Board 1879.1 (2004): 9-16.

[4] http://www.xmdr.org/

[5] Moon, SeokJae, GyeDong Jung, and YoungKeun Choi. "XMDR-DAI Based on GQBP and LQBP for Business Process." Advanced Computer Science and Information Technology. Springer Berlin Heidelberg, 2010. 72-85.

[6] Mell, Peter, and Timothy Grance. "The NIST definition of cloud computing (draft)." NIST special publication 800.145 (2011): 7.

[7] Vassiliadis, Panos, Alkis Simitsis, and Spiros Skiadopoulos. "Conceptual modeling for ETL processes." Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP. ACM, 2002.

[8] Williams, Steve, and Nancy Williams. The profit impact of business intelligence. Morgan Kaufmann, 2010.

[9] Vitt, Elizabeth, Michael Luckevich, and Stacia Misner. Business intelligence. O'Reilly, 2010.

[10] Dayal, Umeshwar, et al. "Data integration flows for business intelligence."Proceedings of the 12th International Conference on Extending Database

Technology: Advances in Database Technology. Acm, 2009.

문 석 재(Moon, Seok Jae)

- 2002년 2월: 광운대학교 전자계산학과(이학사)
- 2004년 2월: 광운대학교 컴퓨터소프트웨어학과(공학석사)
- 2011년 2월: 광운대학교 컴퓨터과학과(공학박사)
- 2006년 ~ 현재: 광운대학교 외래교수
- 2013년 ~ 현재: 전자넷 책임연구원,
- 관심분야: 빅데이터, 클라우드
- E-Mail: msj8086@kw.ac.kr

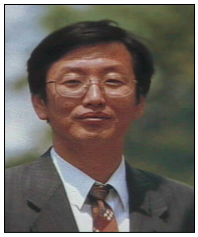
정 계 동(Jeong, Kye Dong)

- 1985년 2월: 광운대학교 전자계산학과(이학사)
- 1992년 2월: 광운대학교 산업정보학과(공학석사)
- 2000년 2월: 광운대학교 컴퓨터과학과(공학박사)
- 1993년 ~ 현재: 광운대학교 교수
- 관심분야: 클라우드 컴퓨팅, 메타데이터
- E-Mail: gdchung@kw.ac.kr

이 종 용(Lee, Jong Yong)

- 1983년 2월 : 한양대학교 원자력공학과(공학사)
- 1988년 2월 : 광운대학교 전자공학과(공학석사)
- 1993년 2월 : 광운대학교 전자공학과(공학박사)
- 2005년 ~ 현재 : 광운대학교 교수
- 관심분야: 자동제어, 로보틱스, 영상인식
- E-Mail: jyonglee@kw.ac.kr

최 영 근(Choi, Young Keun)



- 1980년 2월: 서울대학교 수학교육학과(이학사)
- 1982년 2월: 서울대학교 계산통계학과(이학석사)
- 1989년 2월: 서울대학교 계산통계학과(이학박사)
- 1983년 ~ 현재: 광운대학교 교수
- 관심분야: 객체지향설계, 분산 시스템
- E-Mail: ygchoi@kw.ac.kr