2000년 한국정보처리학회 춘계 학술발표논문집 제7권 제1호

데이터웨어하우스 성장에 따른 개선된 메타프로세스 구현

이동원* 문승진**
*고속도로 정보통신공단
**수원대학교 전자계산학과
*dwlee@hitelecom.co.kr,
**sjmoon@mail.suwon.ac.kr

Enhanced Meta Process Implementation For Growing Data Warehouse

Dong-Won Lee* Seung Jin Moon**
*Highway Telecommunication Corp.
**Dept of Computer Science, Suwon University

요 약

데이터 웨어하우스는 기업의 의사 결정 과정을 향상시킬 수 있게 하는 정보기술이다. 대표적인 정의로는 '기업의 의사결정 과정을 지원하기 위한 주제 중심적이고 통합적이며 시간성을 가지는 비휘발성 자료의 집합 '이다.[1] 즉, 기업들이 보유하고 있는 분산된 대량의 데이터를 추출, 변환, 통합하여 요약된 읽기 전용의 데이터베이스로 구축함으로써, 경영분석이나 기업내의 의사 결정 지원 자료로 주로 활용된다.

데이터 웨어하우스의 경우, 일반사용자는 웨어하우스내에 저장된 데이터를 직접 이용하는 경우가 대부분이다. 따라서, 데이터의 구조와 의미에 대한 일반 사용자의 이해가 필요하게 되었다. 즉, 데이터의 추출 및 정제규칙, 데이터의 통합규칙, 요약알고리즘, 데이터 처리스케쥴 등을 알아야만 한다.

메타데이터는 최소한의 데이터 구조, 데이터의 요약에 사용된 알고리즘, 운영 데이터베이스와 데이터 웨어하우스사이의 대응관계와 같은 정보를 포함하여야 한다.[3] 여기서 변환프로세스에 대한 정보를 데이터의 형식에 대한 정보와 일반적인 데이터들과 차별화하여 메타프로세스라 한다.[5] 메타프로세스는 데이터를 변환하여 데이터 웨어하우스에 적재하는 과정에서 생성되는 메타데이터의 일부로써 데이터 웨어하우스에 통합된 자료들이 어떤 변환과정을 거쳐 생성된 자료인지를 알려주는 변환프로세스에 관한 정보를 제공한다.

본 연구에서는 대부분의 데이터 웨어하우스에서 구현되고 있는 메타데이터들은 데이터 항목의 속성정보를 위주로 한 것이며, 변환 프로세스와 관련된 데이터 관리가 미약하다. 따라서, 데이터 웨어하우스의 메타데이터 중 메타프로세스 정보의 추출 및 관리 시스템을 제안하는 것이다.

1. 서 론

데이터 웨어하우스는 기업의 의사 결정 과정을 향상시킬 수 있게 하는 정보 기술이다. 기업들이 보유하고 있는 대량의 운영데이터를 추출, 변환, 정제하여 대용량의 디스크에 담아 저장한 것으로서, 경영분석이나 마케팅 전략 등과 같은 의사결정과정에서 주로 활용된다.

이와 같은 목적을 가진 데이터 웨어하우스의 구축과정을 살펴보면 먼저, 기업의 다양한 운영 시스템으로부터 데이터를 수집한 후, 특정 주제별로 소스 데이터를 상황에 따라 다양한 방식/관점에서 일관성을 가질 수 있도록 정화하고, 사용자가 쉽게 이해할 수 있도록 통합한 후 데이터 웨어하우스로 전송하여 저장한다. 본 연구에서는 데이터 웨어하우스 구축과정 중에서 데이터 통합부문과 관련하여 운영 데이터로부터 정보 데이터로 추출, 변환, 전송하는 데이터의 통합과정에 대한 메타데이터(Meta Data)의 정보를 표현하고자 한다. 즉, 데이터 웨어하우스의 메타데이터 중 메타프로세스 정보의 추출 및 관리 시스템을 제안하는 것이다.

2장에서는 기존의 메타데이터에 대한 정의 및 중요성과 메타모델의 구축방안에 대해 언급하였고, 3장에서는 Table기반 메타프로세스 관리시스템 구현에 관해 기술하였고, 마지막 4장에서는 결론 및 향후 발전 방향에 관해 기술하였다.

2. 메타데이터의 정의와 메타모델의 구축방안

2.1 메타데이터의 중요성 및 역할

데이터 웨어하우스는 자신이 관리하는 항목들의 목록을 반드시 관리해야 한다. 데이터 웨어하우스의 최종 사용자들은 마치 도서관의 사용자들과 유사하다. 사용자들은 목록을 참조하여 작성한 선택 기준에 따라서 정보를 요구하게 된다. 그들의 요구를 만족시켜주는 프로세스는 데이터 웨어하우스 내의 어디에 정보가 저장되어 있는지 반드시 알고 있어야 한다. 그러므로 데이터 웨어하우스는 자신이 관리하는 정보에 대한 목록 기능을 충족시키는 요소 즉 메타데이터를 반드시 지녀야 한다.

데이터 웨어하우스 구축 과정은 엔지니어링 프로세스이다. 그러므로 이 과정은 발전적인 품질 향상을 위해서 그리고 재생산성을 제공하기 위해 명시적으로 반드시 문서화 되어야 한다.

이 과정의 결과로 발생된 메타데이터는 그러므로 변경 관리의 목적을 위해서 버전관리가 되어야 한다. 예를 들면, 데이터 웨어하우스 목록의 모든 운영계 스키마, 변형 등을 위해 사용된 비즈니스 규칙과 프로세스 규칙들은 반드시 포착되고 버전관리가 되어야 한다.

데이터 웨어하우스를 위한 메타데이터의 설계는 운영계 시스템을 지원하기 위해 데이터베이스의 분석 및 설계와는 매우 상이하다. 운영계 데이터베이스 설계 과정의 초점은 atomic 데이터로 속성이 주어진 정규화된 데이터 모델을 생성하는 것이다.

지금까지의 가장 주요한 관심은 정규화를 사용한 데이터 중복성의 제거였다. 데이터 중복성의 제거의 목적은 데이터 업데이트의 문제를 방지하고 데이터의 일치성을 유지하기 위함이었다. 데이터 웨어하우스를 위한 메타데이터의 설계상의 초점은 종종 상당한 중복성과 함께 수 많은 관계를 사용자에게 제시하는 것이다.

정보 분석가에게 있어서 또 다른 변화는 운영계 시스템이 지니는 현재의 메타데이터에 대한 중요성이다. 대부분의 운영계 시스템은 데이터베이스 및 데이터의 현재의 구조에 대해서 운영된다. 오래된 데이터는 오래된 데이터의 구조와 함께 보존된다. 데이터 웨어하우스 내에서는, historical 데이터를 추출하기 위해서 메타데이터가 사용되어야 한다. 그러므로 운영계 데이터베이스의 메타데이터가 현재부터 제대로 관리되어야 한다.

메타데이터의 중요한 측면은 추출, 정제 및 리엔지니어링 과정을 통해서 소스로부터 데이터 웨어하우스로의 모든 방향으로 맵핑(mapping)을 유지해야 하는 필요성이다. 이러한 맵핑은 다음의 목적을 위해 유지되어야 한다.

- ◆ 데이터 품질의 확인
- ◆ 동기화 및 갱신
- ◆ 통합

데이터 웨어하우스 구축과정 중에서 OLAP을 적용하거나 마이닝 과정을 거치면서도 메타데이터는 생성된다.[10] 그러나 데이터 웨어하우스의 구축과정 중에서 데이터의 이동 및 변환이 가장 많이 발생하는 과정이 바로 데이터 트랜스포메이션이다.

2.2 메타모델의 구축방안

데이타 웨어하우징은 하드웨어나 데이타베이스, 각종 도구 등 많은 구성 요소들을 갖는데 이를 구축하기 위한 주요 기술로써 소스 데이타 처리 기술과 데이타 모델링, 그리고 사용자 지원 기술 등을 들 수 있다. 이들 각 부문을 간략히 정리해본다

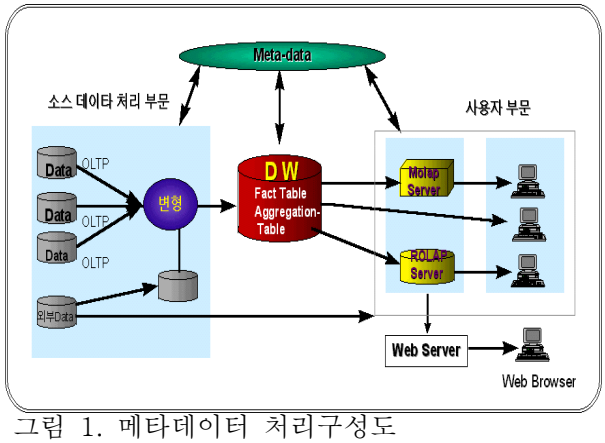

그림 1. 메타데이터 처리구성도

메타데이터베이스는 데이터베이스구축과정에서 개발자들에게 애플리케이션개발에 필요한 정보를 제공하기 때문에 업무메타데이터와 기술메타데이터를 각각 데이터베이스구축대상으로 정의하여 동시에 구축한다.

필요한 데이터를 어떻게 찾을 수 있는가?, 소스데이터는 무엇인가?, 요약데이터가 어떻게 생성되는가?, 데이터를 얻기 위해서 어떤 질의를 해야하는가?, 업무정의와 용어가 지금까지 어떻게 변경되어 왔는가? 등이다. 이러한 질문에 답하기 위해서는 Meta Data Modeling은 다음과 같은 내용을 포함하여야 한다.

- ◆ 정보데이터에 대한 정보
- ◆ 소스데이터에 대한 정보
- ◆ 변형 및 통합규칙
- ◆ 데이터프로세스처리 정보
- ◆ 최종사용자 정보

3. TABLE기반 메타프로세스 관리시스템

데이터 추출부문에는 소스데이터베이스로부터 원하는 데이터를 선택하여 추출하고 이를 적절한 변환과정을 거쳐 타겟데이터베이스에 적재하는 과정으로 볼 수 있다. 본 연구에서는 데이터 웨어하우스 구축과정에서 발생되는 메타 데이터를 이용한 메타프로세스 관리 시스템 구현을 목표로 한다. 메타프로세스가 데이터를 변환하여 데이터 웨어하우스에 적재하는 과정에서 생성되므로 메타프로세스 관리 시스템은 데이터 변환 과정을 시스템에 포함하고 있다.

메타프로세스 관리 시스템에서 데이터 변환 과정은 크게 5 가지 부문으로 구분하여 처리된다.

① Single Table 추출부문
② Multi Table 추출부문
③ 참조 테이블을 이용한 변환 추출부문
④ 시간변이에 따른 Table 변환 부문
⑤ 요약 테이블 생성부문

위의 5가지 유형으로 데이터의 변환 과정에서 발생하는 Meta Process를 정보화하여 사용자에게 유용한 정보를 신속하고 효율적으로 제공하고자 하는 것이다. 다시, 위의 5가지 유형에서 발생된 Meta Process는 6가지로 세분화되어 정보화하게 된다. 첫째는 Table 정보, 둘째는 Column 정보, 셋째는 Filter 정보, 넷째는 변환참조 정보, 다섯째는 시간변이 정보, 여섯째는 요약테이블 정보로 구분하여 표현하게 된다.

즉, 메타프로세스 관리 시스템에는 데이터 웨어하우스내에 존재하는 테이블에 대해 소스데이터베이스내의 어떤 테이블을 통해 데이터를 추출하고, 어떤 Column을 사용하였는지, 데이터의 추출 조건은 무엇인지, 어떤 변환과정을 거쳐 생성되었는지, 시간의 흐름에 따라 데이터의 변경내역, 요약 테이블의 생성과정에 대한 정보를 사용자에게 제공하고자 하는 것이다.

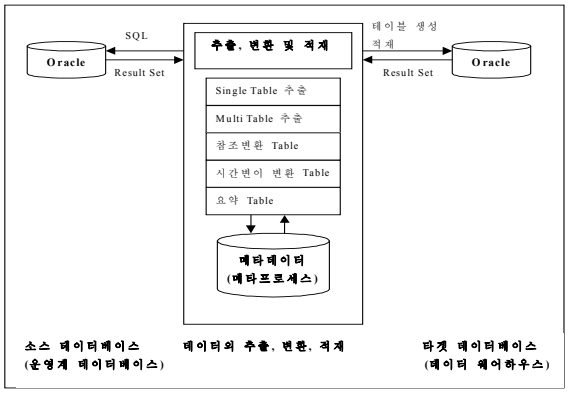

그림 2. Meta Process System Design

4. 결론 및 향후 발전 방향

데이터 웨어하우스는 분산된 대량의 데이터를 통합관리하여 사용자가 요구하는 정보를 빠르고 정확하게 검색하고 이를 의사 결정 지원 자료로 활용하고자 하는 것이다.

현재 대부분의 데이터 웨어하우스 구축도구들은 소스 데이터 및 정보 데이터의 출처나 데이터의 속성정보를 위주로 하는 메타데이터를 지향하고 있다. 따라서, DB 관리자, DB 프로그래머, 분석자 및 일반 사용자들이 진정 알고자 하는 데이터의 변환정보에 대해서는 보여줄 정보가 미약하다. 이를 위해, 본 논문은 데이터의 변환정보

- 필요한 데이터를 어떻게 찾을 수 있을까?
- 소스 데이터는 무엇인가?
- 요약 데이터가 어떻게 생성되는가?
- 데이터를 얻기 위해서 어떤 질의를 해야하는가?
- 요약 데이터의 변환규칙은 무엇인가?
- 시간변이에 따른 DB TABLE 변환 정보는 무엇인가?

등의 정보를 표현하는 메타프로세스 관리 시스템을 구현해 보았다.

향후 연구과제로는 메타데이터 중에서 데이터프로세스처리 정보로서 데이터 처리스케쥴, 데이터 추출이력, 데이터 보관기준 및 이력, 데이터 처리결과 등과 관련된 분야와 최종 사용자 정보로서 가용한 질의어, 데이터 소유자 및 관리자, 데이터 액세스 패턴, 보안 등과 관련된 분야를 세분화하여 정보화 하는 것이다.

5. 참고문헌

[1] W. H. Inmon, "What is a Data Warehouse?", Prism Solutions Tech Topics, vol1.1, no.1, 1995.
[2] W. H. Inmon, "Building The Data Warehouse", 2nd Ed., Wiley, 1996.
[3] Barry Devlin, "Data Warehouse from Architecture to Implementation", Addison Wesley, 1997.
[4] H.Gupta and I. S. Mumick. "Incremental Maintenance of Aggregate and Outerjoin Expressions." Techincal Report, Stanford University, 1999. <http://www-db.stanford.edu/warehousing>
[5] W. J. Labio, D. Quass, B. Adelberg. "Physical Database Design for Data Warehousing." In Proceedings of the International Conference on Data Engineering, Binghamton, UK, April, 1997.
[6] P. O'Neil, D. Quass. "Improved Query Performance with Variant Indexes." In Proceedings of the ACM SIGMOD Conference, Tuscon, Arizona, May, 1997.
[7] V. Harinarayan, A. Rajaraman, J. Ullman. "Implementing Data Cubes Efficiently." In Proceedings of ACM SIGMOD Conference, Montreal, Canada, June 1996.
[8] I. Mumick, D. Quass, B. Mumick. "Maintenance of Data Cubes and Summary Tables in a Warehouse." In Proceedings of the ACM SIGMOD Conference, Tuscon, Arizona, May, 1997.
[9] J. L. Wiener. "What is data warehousing and what is Stanford doing about it?" An overview talk given in the Stanford DB Seminar series, Fall, 1997.
[10] Berson, Smith, "Data Warehousing & Data Mining & OLAP", Mc Graw Hill, 1998.
[11] Debevoise, "The Data Warehouse Method, Design", Prentice Hall PTR, 1999.