데이터품질관리를 위한 어플라이언스 설계

양승연*, 박석천**, 문승식***, 이진희****
*가천대학교 일반대학원 모바일소프트웨어학과
**가천대학교 컴퓨터공학과 정교수(교신저자)
*** ㈜데이터스트림즈 DA본부 상무
**** ㈜데이터스트림즈 DA본부 선임연구원
201240184@gc.gachon.ac.kr

Design of Appliance for Data Quality Management

Seungyeon Yang*, Seok-Cheon Park**, Seung Shig Moon***, Jinhee Lee****
*Dept. of Mobile Software, Gachon University
**Dept. of Computer Engineering, Gachon University
***Managing Director of DataStreams DA Headquater
****Researcher Engineer of DataStreams DA Headquater

요 약

데이터품질관리에 대한 인식과 수요가 증가하고 있다. 그러나 데이터품질관리를 수행하기 위해서는 고려해야 할 사항들이 많아짐에 따라 보다 효과적이고 경제적인 데이터품질관리를 위해 새로운 방안이 모색되고 있다. 데이터품질관리 어플라이언스의 구성은 데이터베이스, 서버, 스토리지, 솔루션으로 이루어져있다. 시스템 구성의 용이성뿐만 아니라 추후 사용자의 관리와 유지보수 체계도 단일화 되어 현재의 시스템보다 사용자의 만족도가 상승할 것으로 판단된다. 본 연구에서는 효율적인 데이터품질관리를 위한 데이터품질관리 어플라이언스의 구성과 체계에 대해 분석하였다.

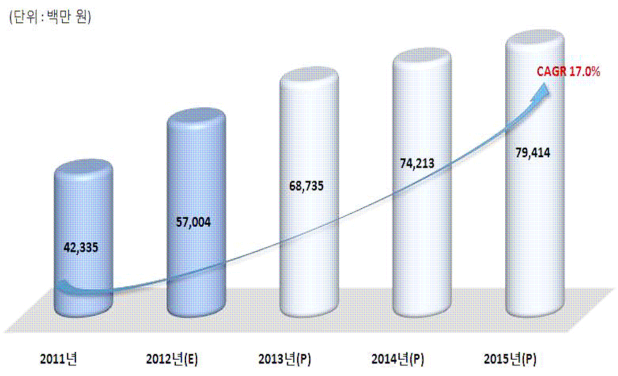
1. 서론

오늘날의 기업들은 변화가 빠르고 경쟁이 치열한 경영 환경에서의 신속한 의사결정을 위해서 경영의 일반적인 기능뿐만 아니라 정보를 효율적이고 효과적으로 획득하고 관리할 수 있는 정보시스템 도입 및 관리에 막대한 자원을 투자하고 있다. 2013년 가트너 선정 10대 전략적 기술 트렌드를 보면 '실용적 분석'이라는 기술이 선정되었다[1]. 이것은 올해에 한해 대두된 이슈가 아니라 '빅데이터'와 같이 데이터의 규모가 커지면서 이에 맞추어 '차세대 분석', '소셜 분석'등 과 같이 2010년 후부터 지속적인 관심의 대상이 되었다.

최근 기업들은 과거 정보 관리의 주요 기능이 확보된 데이터의 양에 의해 좌우되었던 것과 달리, 현재에는 데이터의 양 자체 보다는 응답 속도, 응답 결과의 정확성 등 데이터의 질이 향상되도록 하고 있다. 데이터의 관리의 의미가 확장되어 관리된 데이터로부터 가치 있는 정보를 얻고자 데이터 품질관리를 하게 되었다. 데이터의 품질을 구성하는 많은 영역 중 가장 기본이 되는 영역은 데이터의 정확성 영역이다. 특히 데이터가 단순히 정적으로 저장되어 있는 것이 아니라 끊임없이 변화하는 응용 환경으로 인해 데이터의 정확성 및 일관성 유지 작업은 더욱 어려운 것으로 인식되고 있어서 일회성이 아닌 지속적인 품질관리가 반드시 필요함을 알 수 있다.

데이터 품질 문제로 인한 사회적 비용 및 기업의 손실이 최근 더욱 증가함에 따라 잘못된 데이터를 찾고 관리해주는 데이터 품질관리 도구에 대한 수요 증가를 들 수 있다. 이에 IBM, SAS 등의 대기업은 이미 데이터 통합 솔루션에 품질관리 도구를 포함시켜 시장에 진입하고 있다.

현재 데이터 품질관리 시스템을 도입하여 관리하는 기업과 공공기관이 점차 확대되어 가고 있는 추세이다. 2012년 한국데이터베이스진흥원에서 실시한 DB산업 시장 분석 결과보고서에 따르면 그림1과 같이 데이터품질관리 시장은 2012년에 570억원으로 2011년 대비 34.6% 성장한 것으로 추정된다. 앞으로도 꾸준한 성장을 이룰 것이라고 전망하였다[2].



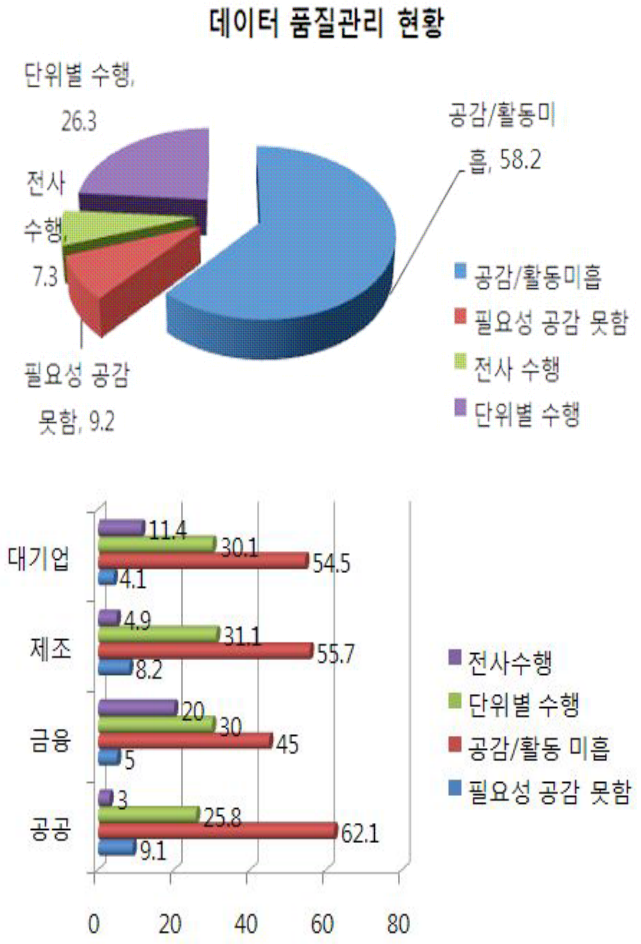
(그림 1) DB산업 시장 분석 결과보고서
출처: 한국데이터베이스진흥원

데이터 품질 시장은 꾸준히 성장하고 있으나 데이터의 품질 관리의 중요성으로 이어지지 못했기 때문에 시장 확대가 부진한 실정이다. 빅데이터라는 바람은 대량의 데이터를 효과적으로 분석하기 위해서 방대하면서도 정확한 데이터로 통합하는 것이 중요하

다는 인식의 전환을 불러왔다.

그러나 통합된 빅데이터의 활용에 있어서 품질이 담보되지 않은 상태의 데이터 오류로 인한 잘못된 정보를 생산하여 활용될 수 있다는 위험에 처해있다. 이렇듯 기업은 빅데이터 활용을 통한 경쟁력 확보를 기대하면서도 데이터 오류로 인한 위험 부담을 동시에 안고 있다.

조사한 바에 따르면 기업들은 여전히 데이터 품질관리에 어려움을 겪고 있는 것으로 나타났다. 한국데이터베이스진흥원에서 2010년에 공공기관 및 기업을 대상으로 조사한 그림 2의 데이터 품질에 대한 인식과 활동 수준설문 평가 보고 자료를 살펴보면 전반적으로 데이터 품질에 대해 필요성은 느끼고 있으나 구체적인 실행은 부족한 것으로 나타났다[3].



(그림 2)데이터 품질관리 성숙수준 조사
출처: 한국데이터베이스진흥원

본 연구에서는 데이터 품질관리를 보다 효율적이고 효과적으로 수행하기 위해 데이터 품질관리 어플라이언스를 제시한다. 하드웨어와 소프트웨어가 통합된 형태의 어플라이언스 시스템의 설계와 데이터 품질관리 어플라이언스 시스템이 데이터 품질관리기술에 대해 미치는 영향에 대해 알아보고자 한다.

2. 관련연구

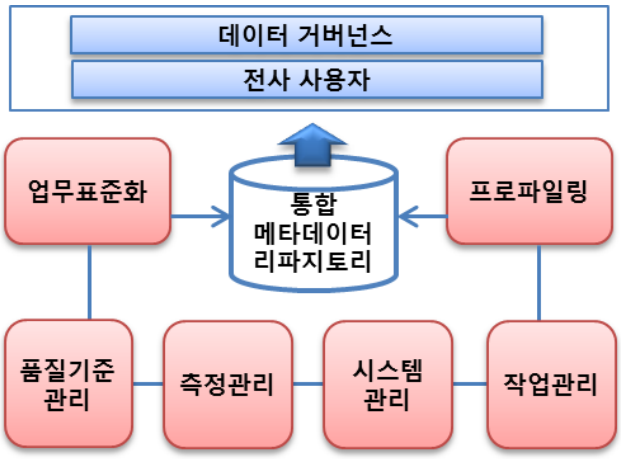
2.1 데이터 품질관리 정의

데이터 품질관리에서 넓은 의미의 데이터는 조직의 전략과 목적을 달성하기 위하여 구축·운영되는 정보시스템과 관련된 모든 자료나 정보를 의미한다. 이때, 데이터는 데이터베이스 내부에 저장되어 있는 데이터 값 이외에 데이터 모델이나 표준 데이터와 같은 구조정보와 문서 형태의 산출물을 포함한다. 실제 품질관리의 대상이 되는 좁은 의미의 데이터는 정보시스템에 저장된 디지털 데이터를 의미한다.

데이터 품질이란 조직의 목적 달성을 위해 관리되는 데이터가 조직 구성원, 고객 등 데이터 이용자의 만족을 충족시킬 수 있는 수준을 의미하고 데이터 품질관리란 조직이 운영하는 정보시스템과 데이터베이스를 활용하는 이용자의 기대를 만족시키기 위해 지속적으로 수행하는 데이터 관리 활동을 의미한다.

최근 기업의 업무가 정보화되면서 업무별 정보시스템 간에 심각한 데이터 중복성과 불일치성의 문제가 대두되고 있다. 예를 들어, 정보시스템 이용자의 의사결정을 효과적으로 지원하기 위해 전사 차원에서 데이터를 통합한 데이터웨어하우스를 운영하는 조직이 증가하고 있으나, 분산된 정보시스템을 통합하고 운영하는 과정에서 정보시스템에 적재된 오류 데이터가 적절히 통제되지 못하고 있다. 이러한 오류 데이터로 인한 잘못된 의사결정으로 피해가 발생하고 있으며 저품질 데이터를 이용한 사용자의 불만도 증가하고 있다[4].

아래 그림 3과 같이 데이터 품질관리 시스템을 통해 관리된 데이터를 사용자에게 제공한다면 오류 데이터로 인한 피해를 줄이고, 조직의 목적을 달성할 수 있을 것이다[5].



(그림 3) 데이터 품질관리 시스템

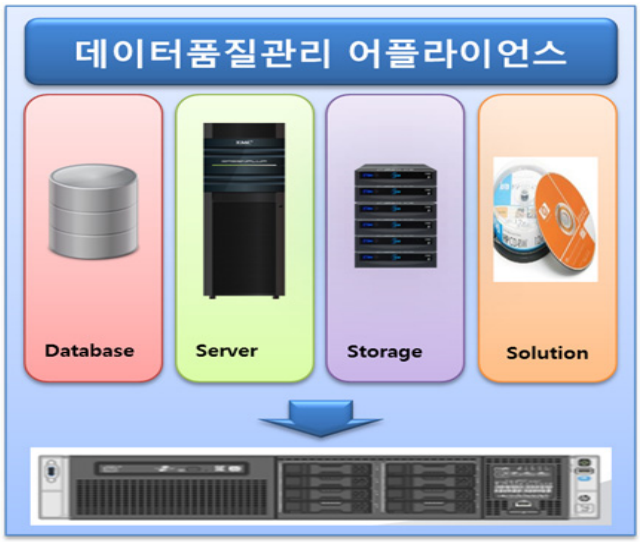
3. 데이터품질관리 어플라이언스 설계

3.1 데이터품질관리 어플라이언스의 구성

어플라이언스는 하드웨어인 서버와 소프트웨어인 솔루션이 포함된 시스템을 말한다. 이와 같은 어플라이언스는 이미 여러 분야에서 이루어 지고 있다. 빅데이터 분석을 위한 오라클의 DB 어플라이언스가 그 예이다. 오라클의 DB 어플라이언스에 오픈소스 배포판 아파치 하둡(ApacheTM HadoopTM), 오라클 NoSQL DB, 하둡을 위한 오라클 데이터 통합 애플리케이션 아답터(Oracle Data Integrator Application Adapter for Hadoop), 하둡을 위한 오라클 로더(Oracle Loader for Hadoop), 오픈 소스 배포판 'R' 등의 솔루션이 포함되어 있다[6].

이와 같은 어플라이언스 시스템은 우발적이며, 잠재적인 물리적 장애 가능성을 낮추고, CPU 코어를 활성화함으로써 비즈니스 요구 증가에 따른 서버 성능의 향상이 가능하고, 솔루션에 요구된 설치환경을 충분히 충족하며, 하드웨어의 성능을 효과적으로 활용할 수 있게 된다. 일반적인 솔루션 설치에도 각 솔루션마다 요구되는 설치 환경의 조건들이 다양하다. 원하는 솔루션의 경우에 맞추어 환경을 재조성하는 경우 이에 따른 비용이 초래되며, 각 하드웨어의 특성과 솔루션의 특징이 맞지 않아 서로 최상의 퍼포먼스를 발휘하지 못하는 경우 있다. 그러나 어플라이언스화하여 이러한 단점을 극복하고 편리함을 최대화 할 수 있다.

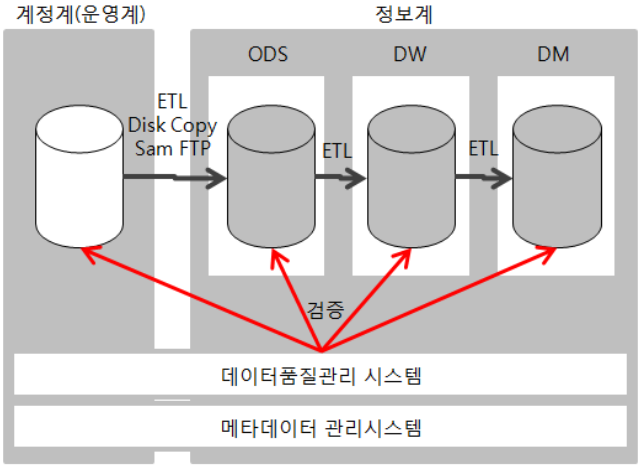
다음 그림 4는 데이터품질관리 어플라이언스의 구성을 나타낸다. 데이터 품질관리 어플라이언스는 데이터베이스, 서버, 스토리지, 솔루션으로 구성되었다.



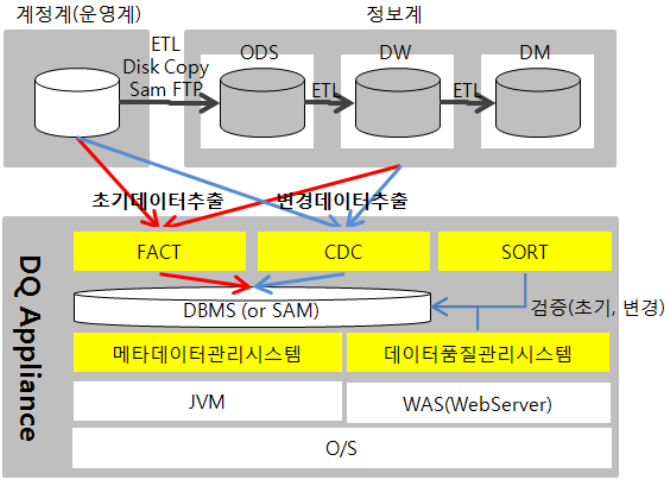
(그림 4) 데이터품질관리 어플라이언스 구성

3.2 기존의 데이터품질관리와 데이터품질관리 어플라이언스의 비교

그림 5는 기존의 데이터품질관리 체계이다. 기존의 데이터베이스품질관리 체계에서는 일방적인 결과만을 제시하기 때문에 데이터품질관리의 한계를 가지고 있었다. 그러나 그림 6에서 보는 바와 같이 데이터품질관리 어플라이언스는 검증대상에 대해서 초기데이터를 추출 후 변경데이터를 추출하는 과정이 가능하여 초기 데이터 추출에 따른 데이터 검증과 변경 데이터에 관한 검증이 가능하여 보다 효율적인 데이터품질관리가 가능하다.



(그림 5) 기존의 데이터품질관리 체계



(그림 6) 데이터품질관리 어플라이언스 체계

데이터품질관리 어플라이언스는 데이터품질관리 기능에 맞춰 스토리지와 DBMS를 최적화하였다. 그렇기 때문에 서버, 스토리지, DBMS를 별도로 도입하여 데이터품질관리시스템을 구축하던 기존의 방식과 달리 별도의 튜닝 작업이 크게 단축되고, 성능이 향상되는 장점이 있다. 경제적인 측면에서 도입 비용이 적고 데이터 모델링의 축소와 DBMS 및 OS의 업그레이드 비용 절감된다. 기존의 데이터품질관리 시스템 보다 성능면에서 뛰어나고, 데이터베이스 아키텍처의 관리가 편리하다. 또한, DBMS나 운영체제, 하드웨어 등 별도의 관리 인력을 최소화 할 수 있어서 관리 및 유지보수가 용이하다.

상용요소 및 오픈 소스의 사용에 따라 하드웨어와 소프트웨어의 비용을 낮추고, 단순한 운용 환경과 용이한 사용 덕분에 설치, 관리, 지원의 비용을 줄일 수 있다.

다음의 표 1은 기존의 데이터품질관리와 데이터품질관리 어플라이언스를 비교한 것이다. 사용자가 데이터품질관리를 위해 각각 DBMS, OS, 서버, 스토리지를 구성하였을 경우 솔루션 설치, 하드웨어 성능, 데이터품질, 데이터검증, 업그레이드, 관리 등 여러 측면에서 어플라이언스 체계보다 효율적이지 못하며, 특히 중요한 데이터품질과 데이터검증의 경우 현격한 기능의 차이를 보이고 있다.

<표 1> 기존의 데이터품질관리와 데이터품질관리 어플라이언스의 비교

구분	기존 데이터품질관리	데이터품질관리 어플라이언스
솔루션설치	기설치된 DB, WAS 등에 설치	전용장비를 이용하여 DQ Appliance 구축
하드웨어성능	타 시스템과 같이 사용하여 최적의 성능을 발휘 못함	H/W, O/S, WAS, DB등을 최적화하여 성능 향상할 수 있는 방안 도출 가능
데이터품질	별도의 추출이 없음	검증대상에 대하여 초기데이터 추출(FACT) 후 변경데이터 추출(CDC)
데이터검증	해당 DB에 직접 접속하여 검증	초기데이터 추출에 따른 데이터 검증 및 변경데이터 검증
업그레이드	각 DBMS, OS, 서버, 스토리지 등 별도 실행	일원화된 유지보수 가능
관리/인력	각 DBMS, OS, 서버, 스토리지 등 별도 관리 인력 필요	관리 인력 최소화

4. 결론

데이터 품질관리란 조직이 운영하는 정보시스템과 데이터베이스를 활용하는 이용자의 기대를 만족시키기 위해 지속적으로 수행하는 데이터 관리 활동을 의미한다. 양질의 데이터를 통한 기업 경쟁력 증진을 위해 기업이 반드시 수행되어야 하는 과업이며, 그 중 가장 기본적이고 핵심적인 요소이다.

데이터의 정확성과 최신성이 확보되면, 이는 빠르고 정확한 정보가 곧 기업의 중요한 역량으로 작용하여 급변하는 환경에 대처할 수 있는 원동력이 될 수 있을 것이다.

본 논문에서는 기존의 데이터품질관리가 가지고 있는 단점과 데이터품질관리 어플라이언스 출현 배경에 대해 기술하였으며 데이터품질관리 어플라이언스 구성과 체계를 분석하고 비교 검토 하였다.

결론적으로 본 논문에서 제안하는 데이터품질관리 어플라이언스는 서버, 스토리지, DBMS를 각각 도입하여 데이터품질관리 시스템을 구축하던 기존 방식을 크게 개선하였다. 또한 튜닝 작업 감소 및 성능 향상, 도입비용 절감, 데이터 모델링의 축소, DBMS 및 OS의 업그레이드 비용 절감, 데이터베이스 아키텍처 관리의 편리성이 증대의 장점을 갖는다. 이것은 기존 데이터품질관리보다 성능 및 유지보수 편리성이 뛰어나기 때문에 앞으로 수요가 증가할 가능성이 높다.

따라서, 이러한 데이터품질관리 어플라이언스가 공공기관 및 기업에 적용되어 보다 효율적인 데이터관리가 될 수 있도록 어플라이언스에 대한 사용자의 인식이 높아져야 할 필요성이 있다.

사사의 글

본 연구는 2013년도 지식경제부의 SW전문인력양성사업의 재원으로 정보통신산업진흥원의 고용계약형 SW석사과정 지원사업(HB301-13-1003)으로부터 지원받아 수행되었습니다.

참고문헌

[1] NIPA, 2013년 주요 IT 트렌드 전망, 2013
[2] 한국데이터베이스진흥원, 2012 DB산업 시장 분석 결과보고서, 2012
[3] 한국데이터베이스진흥, 2010 데이터 품질관리 성숙수준 조사보고서, 2010
[4] 한국데이터베이스진흥원, 데이터품질진단가이드 I
[5] 데이터스트림즈, www.datastreams.co.kr
[6] 정용완, 데이터웨어하우스 어플라이언스 기술 동향 분석, 2012
[7] 고재환, 데이터 품질진단을 위한 자동화도구 개발, 2012
[8] 박주석, 진정한 데이터품질관리의 조건, 2009