빅데이터 환경에서 최적화된 데이터 품질관리 시스템 설계

안하철*, 박석천**, 김종현***
*가천대학교 일반대학원 모바일소프트웨어학과
**가천대학교 컴퓨터공학과 정교수(교신저자)
***위세아이텍 대표이사
e-mail : ahc84@nate.com

Design of Data Quality Management System Optimized in Big Data

Ha-Chul An*, Seok-Cheon Park**, Jung-Hyun Kim***
*Dept of Mobile Software, Gachon University
**Dept of Computer Engineering, Gachon University(Corresponding Author)
***Representative Director, WISEITECH co., ltd

요 약

오늘날 스마트 폰의 보급이 보편화 되면서 모바일 시장이 크게 성장하게 되었다. 또한 그만큼 많은 사용자들이 이용함에 따라 더 많은 양의 콘텐츠를 제공을 해야 하기 때문에 데이터는 점점 증가 할 수밖에 없는 상황이다. 하지만 잘못된 데이터 정보를 마케팅 같은 곳에 활용하여 피해를 보기 때문에 잘못된 데이터와 신뢰성 데이터를 구분을 하여 신뢰성 있는 데이터를 사용자에게 제공해야 한다. 따라서 본 논문에서는 이러한 문제점을 해결하고자 빅데이터에서 추출하는 과정에서 데이터 품질관리를 실시하고 저장된 데이터도 품질관리를 함으로써 신뢰성 있는 데이터를 생성 및 관리 할 수 있도록 하는 빅데이터 환경에서 최적화된 데이터 품질관리 시스템을 설계한다.

Ⅰ. 서 론

미래의 정보기술 가운데 빅데이터(Big Data)는 최고의 이슈로 대두 되고 있다. 스마트 폰의 보급이 보편화 되면서 누구나 스마트폰을 소유하는 시대가 열었고, 이와 더불어 모바일 시장이 크게 성장하게 되었다. 또한 많은 사용자들이 이용함에 따라 더 많은 양의 콘텐츠를 제공을 해야 하기 때문에 그만큼 데이터는 점점 증가 할 수밖에 없는 상황이다.

최근에 소셜네트워크서비스(Social Networks Services), 트위터(Twitter), 카카오스토리(KakaoStory) 등과 같은 새로운 미디어가 등장함에 따라 더 많은 데이터를 생산하게 되었다. 그리고 이곳에 저장된 데이터는 각 사용자의 취미, 성격, 가치를 알 수 있는 중요한 정보로 인식되면서 단순한 데이터가 아닌 가치가 있는 데이터로 되면서 많은 기업의 마케팅에 활용되고 있다.

하지만, 수많은 데이터가 있는 만큼 무분별한 데이터들이 존재한다. 잘못된 정보를 가진 데이터를 마케팅에 활용하였다가 기업은 자금만 소모하는 상황이 발생하고 그 손실로 인하여 빅데이터의 데이터 신뢰성은 떨어질 수밖에 없다. 따라서 쓸모없는 정보, 가치가 없는 정보를 중요한 정보를 가진 데이터로 가공하기 위한 빅데이터 품질관리가 필요하다.

데이터 품질관리(Data Quality Management)는 다양한 형태와 잘못되어 있는 데이터를 신뢰성 있고 가치가 있는 정보로 만들어주는 과정이라고 할 수 있다. 이를 활용하여 신뢰성 있는 데이터 정보를 얻어 활용하면 기업이나 사용자의 만족도는 높아 질 수밖에 없다.

하지만, 그만큼 잘못된 데이터를 찾아내거나 추출하는 것은 쉽지 않다. 그러기 때문에 정확하고 신뢰성 높은 데이터를 얻기 위해서는 효율적인 데이터 품질관리 시스템을 사용해야 한다. 사용자가 원할 때 데이터 품질관리 시스템은 신속하고 정확한 데이터의 정보를 제공하여야 한다.

현재 빅데이터에서 추출하는 과정은 조건 없이 무분별한 데이터를 저장하고 있다. 즉, 데이터를 모두 데이터베이스에 저장하고 있다. 이는 신뢰하지 못한 데이터도 무분별하게 저장되고 있는 것이기 때문에 데이터베이스 공간 효율이 낮아지고 검색도 느려질 수도 있다.

이러한 문제점을 해결하고자 빅데이터에서 추출하는 과정에서 데이터를 품질관리를 하여 무분별하고 정리가 되지 않는 데이터를 신뢰성 있고 높은 가치를 갖는 데이터를 저장할 수 있도록 데이터 품질관리 시스템을 설계하였다.

Ⅱ. 관 련 연 구

2.1 빅데이터 출현배경

빅데이터의 출현 배경은 스마트폰의 확산을 들 수 있으며, 스마트폰 출현으로 가장 가치 있는 개인정보 및 위치정보가 양산되고 있다. 규모로 보면 크지는 않지만 경제적 가치는 최고인 개인정보와 결합되는 빅데이터에 최고의 부가가치를 창출하고 있다[1].

IT혁명에는 인터넷이 세계 경제의 변화를 촉진하고 있으며 그 중 빅데이터는 스마트폰의 핵심자원으로서 세계 경제 변화를 이끌 제 4의 경영자원으로 부상하고 있는 상황이다.

스마트폰 혁명과 소셜 네트워크 효과로 디지털 공간의 데이터 빅뱅이 발생하여 정보에서는 데이터 지식 확보 및 활용 방안을 요구하고, 이에 데이터 활용 시 예산 절감 및 대내외 변화에 대한 신속한 대처, 삶의 질과 정부 신뢰도 향상이 가능해 졌다고 한다.

대표적인 사례로는 그림 1에서 보는 바와 같이 오바마 정부의 필박스(Pillbox) 통한 의료개혁이다. 필박스는 약 검색 서비스로써, 빅데이터를 통하여 수집된 정보의 통계치를 분석하여 연간 5,000만 달러 비용을 절감할 수 있었으며, 독일은 연방 노동기관에서 빅데이터 활용 맞춤형 고용으로 인하여 3년간 백업 유로 비용을 절감했다.

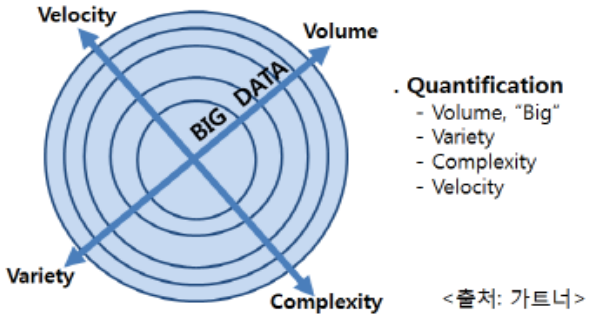
또한 기업에서는 고객 데이터 추적행위 및 수집행위가 증가하고 있는 추세이고, 소셜네트워크서비스의 급격한 확산과 비정형 데이터의 폭증과 정보 수집 가능한 센서 저변의 확대로 매달 수백억 개의 콘텐츠가 페이스북에서 공유되고 있다[1].

(그림 1) 빅데이터 사례(오바마 정부의 필박스)

2.2 빅데이터 정의

빅데이터는 '현재 시스템으로 처리 가능한 범위를 넘어서는 데이터'로 정의된다[7]. 또한, 빅데이터는 페타(Peta: 1015), 엑타(Exa: 1018), 제타(Zeta: 1021)바이트 등 기존의 데이터 단위를 넘어서는 엄청난 양(Volume), 데이터의 생성과 흐름이 매우 빠르게 진행되는 속도(Velocity), 사진, 동영상 등 기존의 구조화된 데이터가 아닌 다양한(Variety) 형태의 정보 등 3가지 속성을 가진다[5].

이처럼, 3가지 속성을 가진 데이터가 '빅데이터'라는 게 대다수 전문가들의 공통된 의견이고 그림 2와 같이 가트너는 3V에 복잡성을 추가해 3V+C로 정의하기도 한다[2].



(그림 2) 빅데이터의 정의

2.3 데이터 품질관리 정의

데이터 품질(Data Quality)이란 조직의 목적 달성을 위해 관리되는 데이터가 조직 구성원, 고객 등 데이터 이용자의 만족을 충족시킬 수 있는 수준을 의미한다.

데이터 품질관리(Data Quality Management)란 조직이 운영하는 정보시스템과 데이터베이스를 활용하는 이용자의 기대를 만족시키기 위해 지속적으로 수행하는 데이터 관리 활동을 의미한다[3].

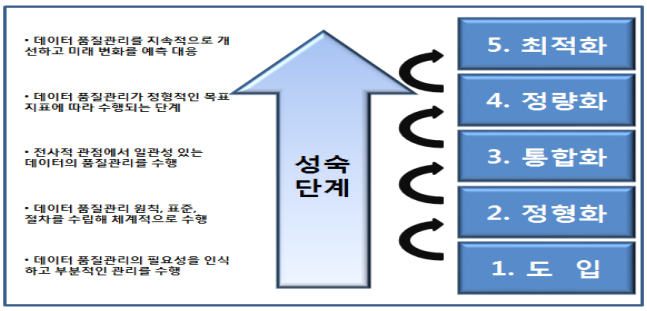
최근 기업의 업무가 정보화되면서 업무별 정보시스템 간에 심각한 데이터 중복성과 불일치성의 문제가 대두되고 있다.

예를 들어, 정보시스템 이용자의 의사결정을 효과적으로 지원하기 위해 전사 차원에서 데이터를 통합한 데이터웨어하우스(DW, Data Warehouse)를 운영하는 조직이 증가하고 있으나, 분산된 정보시스템을 통합하고 운영하는 과정에서 정보시스템에 적재된 오류 데이터가 적절히 통제되지 못하고 있다[3].

이러한 오류 데이터로 인한 잘못된 의사결정으로 피해가 발생하고 있으며 저품질 데이터를 이용한 고객의 불만도 증가하고 있다[3].

2.4 데이터 품질관리 현황

국내 한국데이터베이스진흥원에서는 매년 공공기관 데이터 품질관리 성숙수준을 온라인 설문조사로 진행하여 발표하고 있다[6]. 그림 3에서 보는 것처럼 '도입-정형화-통합화-정량화-최적화'의 1~5 레벨로 측정하고 있으며 2011에 측정된 품질관리 성숙 수준은 1.1 레벨로 조사되었다. 1.1 레벨의 '도입 단계'는 데이터 품질의 문제점과 필요성에 대해 인지하고 부분적인 데이터 품질활동을 시행하는 단계이다[4].

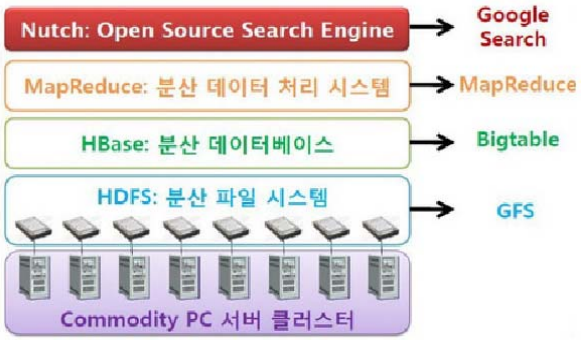


(그림 3) 데이터 품질관리 성숙모형

2.5 빅데이터 처리 기술

빅데이터 분석기법들은 테라바이트 규모의 데이터에 적용되고 있다. 그렇다면 엄청난 규모의 빅 데이터 분석을 수행할 수 있는 인프라 기술에는 하둡(Hadoop), 오프소스 프로젝트 R, NoSQL 등이 있다.

첫 번째로 하둡은 오픈 소스 분산처리기술 프로젝트로, 현재 정형/비정형 빅 데이터 분석에서 가장 선호되는 솔루션이다. 실제로 야후와 페이스북 등에 사용되고 있으며 채택하는 회사가 늘어나고 있다. 하둡의 주요 구성요소는 그림 4처럼 하둡 분산 파일 시스템인 HDFS(Hadoop Distributed File System), HBase, MapReduce의 3가지이다[2].



(그림 4) 하둡 구조 & 대응하는 구글 분산처리기술

Ⅲ. 데이터 품질관리 시스템 설계

3.1 데이터 품질관리 시스템 개요

빅데이터 환경에서 최적화된 데이터 품질관리 시스템을 설계하기 위해 시스템 개요를 구성하였다. 기존에 빅데이터에서 추출한 데이터는 데이터 품질관리 없이 그대로 모든 데이터를 데이터베이스에 저장함으로써 데이터의 신뢰도는 높지 않거나 오류가 발생할 확률이 높아질 수밖에 없는 구조로 되어 있었다.

빅데이터에서 생성된 데이터의 낮은 품질로 인해 많은 문제점이 발생하는 경우가 많았다. 이로 인하여 빅데이터의 신뢰도는 크게 떨어져 활용하는 횟수가 줄어들었으며 기업 입장에서는 빅데이터를 마케팅으로 활용 및 이용하려고 해도 신뢰성이 낮은 데이터로 비용 손실을 보면서까지 빅데이터를 활용하기가 어려울 수밖에 없었다.

따라서 본 논문에서는 이러한 문제점을 해결하고 빅데이터를 활용하는 누구나 신뢰성 있고 가치가 있는 데이터를 활용 할 수 있도록 오류를 줄이고 높은 품질의 데이터를 생성할 수 있도록 데이터 품질 관리 시스템을 설계 하는 것은 목적이다.

3.2 시스템 프로파일링의 오류규칙 분석

데이터 프로파일링의 규칙1은 텍스트, 규칙2는 숫자, 규칙3은 텍스트 + 숫자, 규칙4는 숫자 + 특수문자로 구성되어 있다. 규칙1과 규칙2같은 경우에서는 일반적으로 사용하는 문자나 숫자 등의 공백값, 최소길이, 최대길이, NULL값, 중복값과 같은 오류를 측정할 때 사용한다.

데이터 품질관리 오류규칙 중 규칙3, 규칙4는 기존에 다루기 않았던 새로운 규칙이다. 규칙3은 텍스트+숫자 조합으로 사원번호와 같은 조합으로 이루어진 데이터를 측정하게 되고 규칙4는 전화번호, 주민번호, 핸드폰번호와 같이 숫자 + 특수문자로 이루어져 데이터를 오류 측정을 하게 된다.

기존에는 딱딱 정해진 규칙에만 한정이 되어 전화번호면 전화번호만 되고 핸드폰 번호는 핸드폰 번호만 측정하게 되는 단점이 있었다. 하지만 이번에 새롭게 이루어진 규칙은 사용자가 원하는 자리숫자, 텍스트, 특수문자를 지정하여 효율적인 오류측정을 할 수 있도록 하였다. 본 논문에서 제안하는 시스템의 프로파일링 오류규칙을 분석한 것은 다음 표1과 같다.

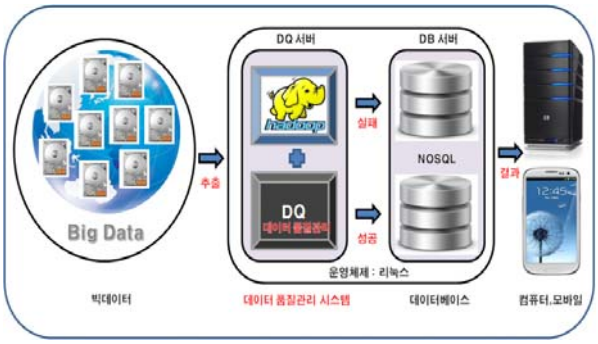
<표 1> 시스템 프로파일링 오류규칙 분석

규칙	프로파일링	진단 및 추가	비고
규칙1	텍스트 (문자)	- 공백 값 허용여부 - 최소최대길이 - Null 값 허용여부 - 중복 값 제거	- ABC - 데이터품질
규칙2	숫자	- 공백 값 허용여부 - 최소최대길이 - Null 값 허용여부 - 중복 값 제거	- 21212515 - 55145665
규칙3	텍스트 + 숫자	- 문자열과 숫자가 포함되어 있는 경우 - 텍스트와 숫자 자리 수 지정	-사원번호 (N001) - 인증번호 (YM23)
규칙4	숫자 + 특수문자	- 숫자와 특수문자가 포함 되어 있는 경우 - 숫자와 특수문자 자리 수 지정	- 전화번호 (010-1111-2222) -주민등록번호 (841111-1212122)

3.3 데이터 품질 관리 시스템 구성도

그림 5는 본 논문에서 제안하는 빅데이터 환경에서 최적화된 데이터 품질관리 시스템의 구성도이다. 빅데이터의 수많은 데이터를 DQ(Data Quality)서버를 통해 추출을 하게 된다. 빅데이터에서 추출된 데이터는 하둡의 맵리듀스 기술을 활용한 데이터 품질 관리시스템을 통하여 데이터의 오류를 측정하게 된다.

데이터 품질관리 시스템을 거친 데이터는 성공한 데이터와 실패한 데이터를 구분을 하여 데이터베이스에 저장하게 되는데 데이터베이스에 저장된 데이터를 사용자가 컴퓨터나 모바일을 통하여 데이터 결과를 볼 수 있도록 구성한다.

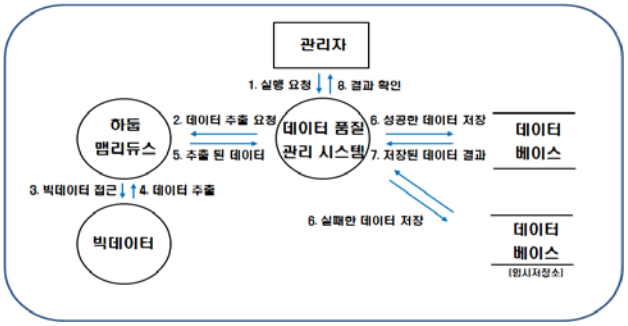


(그림 5) 시스템 구성도

3.4 빅데이터 품질관리 시스템 배경도 및 동작절차

본 논문에서 설계한 데이터 품질관리 시스템 배경도(Context Diagram) 및 동작 절차는 그림 6과 같다.

① 관리자가 데이터 품질관리 시스템에 실행을 요청한다.
② 데이터 품질관리 시스템은 하둡 맵리듀스 기술을 활용하여 데이터 추출을 요청한다.
③ 하둡 맵리듀스를 통하여 빅데이터에 접근한다.
④ 하둡 맵리듀스는 빅데이터에서 데이터를 추출한다.
⑤ 빅데이터에서 추출된 데이터는 데이터 품질관리 시스템에 임시로 저장된다.
⑥ 데이터 품질관리 시스템을 통한 성공한 데이터, 실패한 데이터를 데이터베이스에 저장한다.
⑦ 데이터베이스에 저장 된 결과 데이터를 데이터 품질관리 시스템 전송한다.
⑧ 데이터 품질관리 시스템을 통해 데이터의 결과를 관리자가 확인한다.



(그림 6) 시스템 배경도 및 동작절차

Ⅳ. 결 론

최근 스마트 폰이 보편화되어 모바일 시장의 규모가 광범위해졌으며 이와 함께 많은 콘텐츠가 생성 및 발전함에 따라 데이터의 양이 증가했다. 즉, 빅데이터 환경이 구축되었으며 그 만큼 빅데이터는 중요하고 활용할 기대치가 높아졌다.

그러나, 빅데이터의 데이터 품질이 정확하지 않고 오류가 있으면 데이터를 이용하는 사람은 잘못된 정보로 판단을 하여 피해를 보기 때문에 빅데이터에 대한 신뢰도는 떨어질 수밖에 없다.

따라서, 본 논문에서는 빅데이터에서 데이터를 추출하는 과정과 저장된 데이터 품질관리를 함으로써, 불필요한 데이터, 오류가 있는 데이터, 중복된 데이터를 제거 하여 데이터의 가치를 높이고 신뢰성 있는 데이터를 생산 및 활용하도록 제안했다.

향후, 빅데이터 품질관리 시스템을 구현하고 테스트를 진행하여 생성된 결과를 바탕으로 제안 시스템을 수정 보완할 예정이다.

사사의 글

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

참고문헌

[1] 최 성, 우성구 “빅데이터 정의, 활용 및 동향”, 2012
[2] 김정숙 "빅 데이터 활용과 관련기술 고찰", 2012
[3] 한국데이터베이진흥센터 “데이터 품질 가이드라인”, 2011
[4] 한드림 “데이터 품질이 공공기관의 서비스 품질에 미치는 영향에 대한 사례연구”, 2011
[5] 김병곤 “빅데이터 기반 기술을 활용한 분산 처리 및 실시간 처리 방안“, 2012
[6] 문성은 “메타데이터와 연계한 데이터품질관리의 경제적 효과 분석 및 사례 연구”, 2013
[7] 정성우 “빅데이터 개요와 관련 기술, 그리고 오라클의 지원 전략”, 2012
[8] 이미영 “빅데이터 분석을 위한 빅데이터 처리 기술 동향”, 2012