논문 19-44-05-22 The Journal of Korean Institute of Communications and Information Sciences '19-05 Vol.44 No.05
https://doi.org/10.7840/kics.2019.44.5.957

Bigdata 시대에서 DataWarehouse 대응방안: DataWarehouse 중심으로

김 재 형◆, 김 석°, 장 동 원°

DataWarehouse Countermeasures in the Bigdata Era: Focused on DataWarehouse

Chae-Hyeong Kim◆, Seog Kim°, Dong-Won Jang°

요 약

최근 4차 산업혁명과 더불어 다양한 IOT기반의 데이터가 대량으로 증가하고 있는 추세이다. 이러한 대량의 데이터는 기업에서 경쟁력 확보 방안과 정확한 의사결정을 지원하기 위하여 DW 시스템을 구축하여 업무에 적극적으로 활용하고 있다. 그러나 매년 증가하는 대용량의 데이터와 다양한 비정형 데이터 분석의 필요성이 더욱 확대되는 상황에서 기존 DW 시스템에서는 분석의 한계에 봉착하게 되었다. 본 연구에서는 기존 DW 시스템의 문제점을 파악하고, 대용량 데이터와 비정형 데이터 분석을 DW 시스템에서 처리할 수 있는 방안을 모색하여 저장 아키텍처, DW 및 빅데이터 아키텍처로 모듈화 하여 확장된 DW 시스템에서 비정형 데이터 분석을 위한 방안을 제시하고자 한다.

Key Words : Data Warehouse, Bigdata, Data Lake, Hadoop, Data Mining, OLAP

ABSTRACT

In recent 4th industrial revolution, various IOT-based data are increasing in volume. This large amount of data has been used in business by building a DW system to support competitiveness and accurate decision making in enterprises. However, the need to analyze large amounts of data and various unstructured data, which are increasing every year, has been further enlarged, resulting in limitations of the existing DW system. In this study, we identify the problems of existing DW system, find ways to process large amount of data and unstructured data in DW system, modularize into DW architecture, DW and big data architecture and analyze unstructured data in extended DW system This paper proposes a solution for this problem.

Ⅰ. 서 론

1999년 이후 국내 기업의 의사결정 지원을 위해 도입된 DW(Data Warehouse) 시스템은 최근까지 많은 발전을 거듭하고 있지만, 증가하는 데이터양과 다양한 비즈니스 분석의 필요성과 그에 대한 결과물을 명확하게 제시하지 못하고 있다. 기존 전통적인 DW 시스템으로는 정형화된 데이터의 분석과 사용자 중심보다 IT적인 관점에서 시스템이 구축되어 이런 요구에 부응하기에는 아직은 미흡하다. 그 결과 최근 대량의 비정형 데이터를 분석하기 위해 이미 다수의 기업들이 빅데이터를 도입하여 운영을 하고 있거나, 신규로 빅

◆ First Author : Soongsil University Graduate of IT Policy Management, camelz@naver.com,정회원
° Corresponding Author : Soongsil University Graduate of IT Policy Management, nature.kim@gmail.com, 학생회원; jjangdongwon@hanmail.net, 정회원
논문번호 : 210902-452-0-SE, Received February 8, 2019; Revised March 11, 2019; Accepted March 15, 2019

데이터 도입을 적극 고려하고 있다. 또한, 다양한 매체를 통한 정형, 비정형의 데이터가 매일 수백, 수천 이상의 방대한 자료가 발생되고 있는 상황에서 경쟁기업간의 선두다툼에서 손 놓고 눈치만 볼 상황은 아니라는 점이 반영 되었던 것이다. 그러나, 최근 빅데이터를 도입한 몇몇 기업들은 또 다른 딜레마에 빠져드는 향상을 보이고 있다. 이미 이전에 도입된 DW 시스템과 빅데이터 시스템간의 업무처리 방향에 적지 않은 혼선과 의사결정과정에서 즉시성에 대한 두 시스템의 평가가 양분되고 있다는 점이다.

최근 정보통신기술의 발전이 가져온 엄청난 변화는 이제 산업분야를 넘어 우리 일상생활에까지 영향을 미치고 있다. 이런 ICT 기술을 접목한 수많은 기기들이 하루가 멀다 하고 새롭게 탄생하고 있고 각종 기기에서 발생되는 수많은 데이터와 스마트 시대에서 SNS, 모바일, IoT 등 다양한 매체에서 생성되는 데이터의 분석이 이제 우리 인류의 삶의 변화를 가져올 만큼 중요한 시대가 되었다. 데이터 분석이 새로운 패러다임을 가져오게 된 것이다. 과거 전통적인 DW를 운영하는 기업도 새로운 패러다임의 변화를 수용하고 이제 DW 시스템을 확장하거나 다른 방안을 모색하여 다양한 요구에 부응해야 할 시기인 것이다. 그렇다고 기존 DW 시스템을 포기하고, 빅데이터만 새롭게 구축하여 운영한다고 해서 모든 문제가 해결되지 않는다. 비정형 데이터도 중요하지만, 기업의 내부에서 발생되는 데이터는 대부분 정형 데이터로 회사 정책에 큰 영향을 주는 양질의 데이터이다. 결국 기존 DW 시스템을 최대한 활용하면서, 빅데이터 분석을 원활하게 할 수 있어야 가장 효과적이고, 합리적인 방향이라고 할 수 있다. 그리고 비용적인 측면과 효율성 부분도 감안해야 한다. 이러한 사전 고려사항들을 충분히 검토한 후에 빅데이터 분석을 위한 시스템을 구축하거나, 고도화 전략을 세워야 할 것이다.

Ⅱ. 관련연구

2.1 DW 정의

DW란 정보(data)와 창고(warehouse)의 합성어로 기업의 정보 자산을 효율적으로 활용하기 위한 시스템이다. 기업의 전략적 관점에서 최적의 의사 결정을 지원하기 위해 정보의 집합으로 DW는 방대한 조직 내에서 분산 운영되는 각각의 데이터베이스 관리 시스템들을 통합하여 조정·관리하며, 기업경영 전반의 의사 결정 시스템을 위한 주제 지향적이고, 다양한 원천 데이터를 수집, 가공하여 효율성 있는 정보 제공을 목표로 구축된 시스템이다. DW의 또 다른 정의를 살펴보면 의사 결정을 지원하기 위한 데이터를 주제(업무영역)중심적으로 구성하고, 다양한 데이터 통합과 분석을 위한 일정 기간의 시간분포를 구성하는 시계열적 자료생성, 그리고 대용량의 데이터를 일관적으로 보관하는 비휘발성으로 모아 놓은 것이라고 정의한다[6]. DW의 특징을 정리하면 <표 1>과 같다.

초기 전통적인 DW는 트랜잭션 중심의 운영계 시스템에서 필요한 정보를 취합하여 각 주제별로 마트를 구축한 이른바 중앙집중식 데이터베이스 구조로 기업의 핵심 의사 결정 지원 시스템을 수행해 왔다. 아래 <표 2>는 DW 구축전과 구축 이후의 효과에 대한 내용이다. 기본적으로 트랜잭션 위주의 시스템 운영을 효율성과 데이터 가공의 시간 단축이라는 장점

표 1. DW의 특징
Table 1. Features of DW.

	Description
Subject	업무영역별로 특정지어 하나의 주제 중심으로 구성
Data Integration	다양하게 분포된 원천시스템의 데이터 수집, 통합적재
Time Variation	데이터의 추출 주기를 년, 월, 일 등 일정기간의 시점을 기준으로 데이터 수집
Nonvolatile	적재된 데이터의 일관된 분석을 위한 변경이 발생하지 않도록 유지

표 2. DW 구축 효과
Table 2. DW building effect.

	Before	After
Integration	- 업무별 기준에의한 상이한 정보로 통합 정보 제공 제한	- 전사적 관점의 통합 정보제공
Efficiency	- 마트별 정보가공처리 중복, - 데이터 정합성 제한	- 마트간 정보 신뢰성 향상 및 유지, 관리환경 최적화
Timeliness	- 데이터 가공 소요시간의 증가로 정보제공 적시성 제한	- 정보 수집 및 가공시간 단축으로 정보제공 적시성 개선
Reliability	- 데이터 표준 및 품질관리 체계 부재로 데이터 신뢰도 저하	- 전사적 정보관리체계 일원화 - 데이터 표준화 및 정보신뢰성 향상
Rapidness	- 일회성 보고를 위한 데이터 중복 가공으로 정보제공 응대시간 장기화	- 관련 마트구축으로 사용자 요구 응대시간 단축

을 바탕으로 기업의 의사 결정을 위한 정보 제공에 큰 역할을 하였다. 그러나 최근 4차 산업혁명과 방대한 데이터의 증가로 인하여 이에 대한 분석의 필요성이 절실히 요구되고 있다. 또한, 각종 디지털기기와 SNS, 스마트미디어 등에서 발생되는 수많은 정보를 실시간으로 분석하여 기업운영에 필요한 중요 분석 자료로 활용하려는 기업들이 늘어남에 따라 기존 전통적인 DW 시스템으로는 한계를 느낄 수밖에 없다. 이에 대한 해결방안으로 수년간 각 기업들은 기존 DW를 발전시키는 차세대 DW 시스템을 구축하여 고도화 단계를 수행해 왔다. 대표적인 방안으로 BI 솔루션을 도입하여 다양한 분석과 시각화 지원으로 최근까지 기업의 의사 결정 지원 시스템에서 중요한 역할을 하고 있다.

2.2 DW 고도화

전통적인 DW에서는 다양한 업무영역의 사용자들이 방대한 데이터를 좀 더 쉽게 가공하여 업무에 활용하기 위한 방편으로 시스템을 구축하여 운영해 왔다. 초기 수많은 트랜잭션이 발생되는 기능 중심의 운영계 시스템에서 업무별 주제를 기준으로 중앙집중식의 데이터 저장소 개념을 적용한 것이다.

당시로서는 이러한 데이터의 집합만으로도 기업의 업무변화에 상당히 많은 기여를 했으며, 방대한 데이터를 적절한 분산과 통합으로 작업의 능률을 향상시킬 수 있었다. 최근에는 DW 시스템의 발전으로 다양한 변화를 가져왔고 기존 DW를 운영하던 기업들도 점차 차세대 DW 시스템을 구축하기 위한 고도화 프로젝트를 추진하고 있다. 즉, DW의 패러다임이 새롭게 변하고 있다. 기업에서 의사결정권자는 유용하고 적절한 시점에 가장 신뢰할 수 있는 정보를 제공받기를 원하고 있다. 또한 고부가가치를 실현하기 위해서 다양한 비즈니스 분석이 필요하고 기업의 특성이나 미래예측 등을 위한 트렌드를 분석할 필요성을 절실히 느끼고 있다.

4차 산업혁명과 맞물려 다양한 정보의 분석과 대용량의 데이터 처리가 필요한 시기에 기존의 DW 시스템으로는 한계에 봉착했다고 볼 수 있다. 하지만, 정형 데이터처리나 기존 운영 업무에서 탁월한 성과를 보인 DW 시스템이 낙후된 것이 아니라, 단지 대용량의 데이터 처리나 실시간 데이터분석 등 최근 빅데이터 환경만 제공하는 기준에서는 기존 DW 시스템으로는 감당하기 어렵다. 차세대 DW 시스템에서는 이러한 문제점을 해결하기 위해서 하드웨어나 소프트웨어 업체들이 결합하여 DW 어플라이언스 제품을 출시하고 대용량의 데이터 분석과 빠른 처리속도 등의 목적으로 제안하고 있다. 또한, BI 솔루션 도입도 차세대 DW 시스템 구축의 가장 큰 특징이다. DB 분석 솔루션은 DW의 생성과 관리, BI, OLAP, Data Mining, 분석 등 의사결정을 지원하기 위한 도구를 총칭한다[2]. 빅데이터 시대를 맞이하게 된 현재 DW 시스템의 운명은 끝이 아닌 새롭게 다시 정비하고, 보완하여 기존 체계화된 정형데이터 분석과 다양한 비정형 데이터 분석과 대용량 데이터 및 실시간 처리를 병행할 수 있는 DW 시스템으로 전환되어야 한다. 그리고 기업의 요구사항에 잘 부응하는 맞춤형 DW 시스템으로 확장되어야 한다.

2.3 빅데이터 정의

빅데이터에 대한 정의는 해외 주요 기관이나 연구에서 다양하게 정의되고 있다. 보는 관점에 따라 약간의 차이점이 있을 뿐 대부분 유사하다. 즉, "데이터가 매우 크고(대용량), 복잡하며, 다양한 종류의 데이터를 빠르게 처리할 수 있는 기술의 집합체"를 빅데이터라고 정의하였다. 빅데이터의 특징은 일반적으로 3V로 설명을 한다. 3V는 규모(Volume), 속도(Velocity), 다양성(Variety)의 특징을 가지고 있다[3]. 이런 3가지 요소로 기존 DW 시스템과는 다른 시스템으로 분류된다. DW 시스템은 분석을 위하여 특정 목적별로 데이터를 모아 처리한다. 이에 반해 빅데이터는 다양한 데이터 흐름을 연속적으로 빠르게 처리할 수 있는 시스템이라고 할 수 있다. 최근엔 빅데이터 속성이 기존 3V에서 좀 더 확장된 개념으로 4V, 5V 등 업무특성이나 관련 기술의 영향을 받아 새로운 속성들이 추가되고 있다. 정확성(Veracity), 가변성(Variability), 시각화(Visualization) 등 추가적인 속성들을 상황에 따라 적용하려고 하지만, 기본 속성인 3V 특징은 변하지 않고 있다.

표 3. 빅데이터 3가지 특징
Table 3. BigData 3 Features.

	Description
규 모 (Volume)	데이터의 크기, 대용량의 데이터 데이터 단위가 TB, PB, EB 등의 규모
속 도 (Velocity)	데이터 처리 속도를 나타냄 Batch에서 실시간 처리로 이동
다양성 (Variety)	데이터의 다양성을 의미 정형데이터, 비정형데이터

2.4 빅데이터 분석

빅데이터의 분석은 데이터에 대한 3V 속성의 출연

표 4. 빅데이터 분석의 4대 요소
Table 4. The Four Elements of Big Data Analysis.

	Description
분석준비 및 성숙도 진단	현재 내부의 분석 수준파악
분석 전략 계획수립	분석에 통한 구체화
분석 관련 지식 내재화	업무적용사례 조사
분석적 사고 내재화	분석 역량 강화, 중요성 인식

으로 새로운 인식이 대두 되었고, 급변하는 상황에서 기업의 경쟁력 확보방안과 Risk 개선, 그리고 다양한 분야가 융합된 새로운 통찰력을 제시한다. 빅데이터 분석을 위한 각 사전 계획 단계의 요소는 <표 4>과 같이 4가지로 요약할 수 있다.

빅데이터 분석의 계획이 설정되면 이후 활용을 위한 각 단계별 프로세스를 구성하고, 프로세스별 기술을 접목하여 체계화된 구조가 빅데이터 플랫폼이다. 빅데이터 분석은 기존의 통계적 또는 수학적 분석과 방법이나 기법에서 차이가 있다. 전통적인 통계분석은 Sample을 기준으로 모수를 추정하는 방식으로 두 개의 집단을 비교 분석하는 방식이고, 빅데이터 분석은 기본적으로 모수가 존재하고 있고, 해당 모수에서 특정의 패턴과 규칙을 찾는 방식이라고 할 수 있다. 일반적인 통계분석의 방식으로는 정규분포, 표준편차, 샘플 및 모집단, 그리고 전통적인 통계적 분석의 기법이 사용되고 있다. 통계적 분석의 기준은 주로 수치형의 데이터에서 확률을 기반으로 특정현상의 추정 및 예측을 하는 기법이다. 빅데이터 분석에서는 수치형기반보다 데이터의 패턴을 찾고, 추이를 분석하는 방식으로 처리된다.

2.5 Data Lake 정의

Data Lake는 정형 데이터 및 비정형 데이터를 원시형태 그대로 저장하고 있다가 필요에 따라 수시로 해당 데이터를 분석, 가공할 수 있도록 지원하는 대용량의 데이터 저장 기술을 의미한다. DW 시스템은 정형화된 주제영역을 최적화를 통해 확실히 신뢰할 수 있는 정보를 제공하는 시스템이라면, Data Lake는 구체적이고 정확한 결과의 제공보다 대략적인 추이를 분석하는 패턴처리에 더 가깝다고 할 수 있다. 실제 업무의 활용에 있어서도 두 시스템의 차이는 분명하다. 즉, 정확한 결과를 제공해야 하는 수치적인 분석이나, Key 맵핑이 가능한 업무는 DW 시스템이 더 적합하고, 추세적인 부분이나 전망 등 데이터 흐름을 기준으로 분석을 할 경우에는 Data Lake가 적합하다. Data Lake가 등장하게 된 배경은 DW 시스템의 한계를 극복하기 위한 방안이다. 즉, 대량 데이터 및 비정형 데이터를 처리하고 장기간의 원시데이터를 보존할 방법이 필요하게 된 것이다. Data Lake 아키텍처는 빅데이터의 Hadoop 플랫폼을 기반으로 정형, 비정형 데이터의 빠른 로드와 저장, 그리고 저렴한 구축비용 장점이다. 기존 DW 시스템의 Data Mart와는 다른 구조의 아키텍처로 설계되었다. 그러나 Data Lake도 여러 문제점을 가지고 있다. 그중에서 One-way Data Lake라는 Garbage Dump 문제를 가지고 있다[9].

2.6 정형, 비정형 데이터 정의

정형 데이터란 1차 가공된 데이터로 원시 데이터에서 분석에 사용할 수 있도록 체계화된 데이터를 의미한다. 일반적인 관계형 데이터베이스에서 주로 사용이 되고, 고정된 필드 구성으로 저장된다. 정형 반대의 의미로 비정형 데이터가 존재하는데, 비정형 데이터란 최근 분석의 필요성이 부각되고 있는 동영상 정보나 음성 정보, 그리고 SNS에서 Text 등 가공되지 않은 Raw Data를 의미한다. 정형과 비정형 데이터의 공존은 과거 기능적인 프로세스 방식에서 업무 요구에 맞게 데이터를 가공하고 분류하던 시스템에서 저장된 정형 데이터와 정형화될 수 없는 자료, 즉, 일상적인 자연어와 같은 의미의 정보를 새롭게 분석하기 위해 버려지거나 무시돼 왔던 자료를 보관하기 시작한 시기부터이다. 이러한 정형과 비정형 자료는 다양한 매체를 통하여 폭발적인 증가추세를 보이고 있다.

기존 DW 시스템이 정형화된 데이터분석에 강점이 있다면 빅데이터 시스템에서는 비정형 데이터를 수집하고 가공하는 다양한 Ecosystem 등장으로 다양한 데이터를 원활하게 분석 할 수 있다. 데이터의 증가가 단지 비정형 데이터의 포함으로 방대해진 것은 아니고 정형 데이터도 많이 발생하고 있다. 기업에서는 여전히 정형 데이터 가 비정형 데이터 보다 더 의미 있는 자료로 분류하고 있다. 기업에서 정형 데이터는 대략 20~30% 정도이고, 나머지 70~80%가 비정형 데이터로 분류되고 있다. 비정형 데이터가 차지하는 비중이 높아짐에 따라 정형 데이터로는 대략 2~30%의 정보만 획득하고 나머지는 비정형 데이터를 분석해야만 실질적인 고객의 서비스나 기업의 경쟁력 확보에 중요한 자산이 될 수 있다.

Ⅲ. 기존 시스템의 문제점

3.1 빅데이터 시대의 DW 시스템 문제

앞장의 관련 내용에서 언급했듯이 전통적인 DW

시스템에 대한 문제점이 존재하고, 그에 대한 해결방안으로 차세대 DW 시스템을 구축하게 된 배경을 설명하였다. 차세대 DW 시스템의 가장 큰 변화는 BI 솔루션을 도입하여 정형 데이터 분석뿐만 아니라, 비정형 데이터를 분석할 수 있는 토대를 마련했다는 것이다. BI 솔루션을 이용한 다양한 데이터 분석이 가져온 특징은 기업의 의사결정 과정을 빠르고, 좀 더 정확한 정보를 획득할 수 있다는 장점이 있었다. 하지만, DW 시스템이 모든 요구사항을 다 수용할 정도로 다재다능한 시스템이라고 할 수는 없다. 데이터 분석이라는 주제에서 다양하고, 방대한 원천 자료를 정확하게 분석한다는 것이 현실적으로나 기계 중심적인 상황에서 쉽지 않은 문제이기 때문이다. 빅데이터는 외부 비정형 데이터 분석을 기준으로 거론되고 있지만, 기업 경영 측면에서 중요한 핵심정보는 내부 정형데이터에서 결정된다고 할 수 있다. 잘 정제된 정형 데이터 분석 결과들과 다양한 형태의 비정형 데이터를 결합하여 통합된 새로운 분석결과를 도출하는 시스템이 현시점에서는 더욱 절실히 필요하다. 그만큼 빅데이터 시대에서도 기업내부의 정형 데이터를 분석하는 DW 시스템이 중요하다고 할 수 있다[4].

기존 DW 시스템의 구축과정은 상당히 복잡하고 많은 어려움이 존재 했다. 비지니스 프로세스를 분석하고, 단계별로 아키텍처를 구성하여 논리적, 물리적 데이터 모델링 과정과 기술적인 결합으로 최종 완성된 DW 시스템을 구축한다. 이후 구축된 DW 시스템을 기반으로 각각의 업무 프로세스에 적용하여 운용하게 되는데, 문제는 이런 일련의 프로세스가 이미 정형화된 구조라는 것이다. 즉, DW 시스템의 일반적인 처리 프로세스가 정형화된 데이터를 기준으로 모델링 된다는 것이 첫 번째 문제점이다. 비정형 분석이 필요하다는 것이다.

두 번째 문제점으로는 전통적인 DW 시스템의 H/W적인 아키텍처 문제를 들 수 있다. 최근 다양한 매체에서 발생되는 데이터와 대량의 데이터 처리가 필요한 시점에서 기존 DW 시스템의 H/W적인 문제가 큰 이슈로 부각되고 있다. 단순히 데이터의 증가만이 문제만은 아니다. 데이터의 종류 또한 다양해서 빠른 분석과 실시간의 분석이 절실히 요구된다는 점이 H/W의 성능이나 기타 주변 인프라가 뒷받침되어야 한다.

표 5. DW 시스템이 당면한 문제점
Table 5. Problems facing the DW system.

	Description
unstructured data analysis	다양한 비정형 데이터분석 필요
Bigdata Analysis	대용량 데이터 처리를 위한 수집과 저장
Real-time processing	실시간 분석의 필요

3.2 데이터 분석의 문제

최근 DW 시스템의 최대 장점은 수년간 고도화를 거듭하며 사용자 요구에 최적화되어 있다는 점이다. 물론, 데이터의 증가량에 비해 DW 시스템의 처리용량은 한계에 봉착했고, 시스템 증설 비용이나 유지보수 비용이 기하급수적으로 늘어나는 문제에 직면해 있다. 또한, 전통적인 DW시스템 아키텍처가 갖고 있는 구조적인 한계 때문에 비정형 데이터나 다양한 경로로 유입되는 자료의 분석은 엄두도 내지 못하는 실정이다. 데이터가 과거에 비해 요즘 생성되는 데이터의 크기가 크고 복잡한 형태를 지니고 있어 숨겨진 의미를 찾는 것이 어렵다는 것이다[5]. 결국 각기 시스템의 장점만을 위한 새로운 DW 패러다임의 변화를 가져오게 되었다. 즉 빠른 분석을 지원하고, 고품질의 데이터로 정제하여 정형 및 비정형 데이터 분석을 DW 시스템에 포함시킨 하이브리드 DW 시스템을 구축하는 것이다.

DW 시스템에서는 정형 분석을 담당하고, 비정형 데이터는 Hadoop 기반으로 빅데이터 플랫폼을 추가하여 구축하는 방식이다. DW와 빅데이터가 유기적으로 연동되어 늘어나는 대용량의 데이터처리와 다양한 분석이 요구되는 비정형 데이터를 효과적으로 활용하기 위한 방안으로 기존 DW 시스템의 문제점과 한계를 빅데이터를 통해 극복하자는 것이다. Hadoop과 RDBMS를 결합한 통합 시스템의 장점은 데이터 처리 시간의 단축으로 병목현상이 자주 발생되던 DW 시스템의 H/W 부하를 감소시키고, 비정형 데이터처리에서 정형 데이터로 정제하여 데이터의 품질을 향상시킬 수 있는 장점이 있다. 또한, 증가하는 데이터의 저장과 관리비용을 Hadoop 시스템으로 구축하여 비용절감 효과를 얻을 수 있는 장점이 있다. 또한, 비정형 데이터와 정형데이터를 결합한 분석이 좀 더 유용한 비즈니스 성과를 얻을 수 있다[8].

3.3 DW 시스템과 빅데이터 병행 사례

최근 DW 시스템에서 빅데이터의 분석이 필요함에 따라 다양한 적용 사례들을 발표하고 있다. 아래는

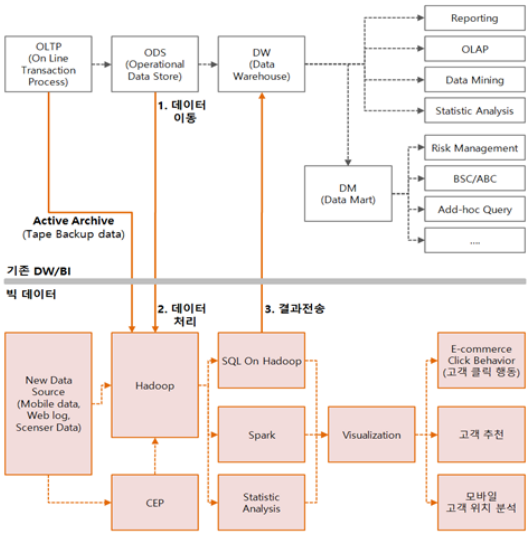

그림 1. 하이브리드 DW 시스템 구성도
Fig. 1. Hybrid DW system configuration diagram.

DW 시스템을 운용하면서 빅데이터를 도입하여 각각 두 개의 시스템에서 정형데이터 분석과 비정형 데이터 분석을 병행으로 처리하는 시스템 구성도이다[1]. 기본적인 특징은 기존 DW 시스템을 운영하면서, 비정형 데이터 처리를 위하여 빅데이터를 도입, 빅데이터의 분석결과를 DW 시스템과 공유하여 분석하는 구조이다. 또한, BI 솔루션의 처리는 각각의 분석결과를 바탕으로 분산 처리하는 구조이다. 해당 제안의 문제는 병행의 구조로 빅데이터와 DW 시스템간의 데이터 통합에서 문제가 있고, 최종 분석이 정형, 비정형으로 이원화 된다는 점이다. 그러나 DW 시스템의 확장성을 고려한 많은 제안들도 유사한 구조로 Hadoop 플랫폼 도입을 적극 활용한다는 점에서 큰 효과를 거두어 왔다.

TDWI(The Data Warehouse Institute) 조사에 의하면 <그림 2> 와 같이 Hadoop(HDFS)이 DW를 대체하기보다 보완하는 역할을 한다고 설문을 통해 제시하였다[7].

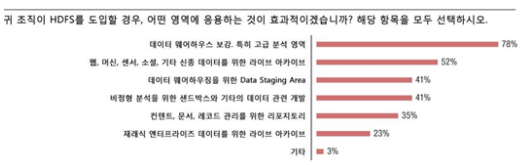

그림 2. HDFS를 도입할 경우 효과성 설문
Fig. 2. When HDFS is introduced, the effectiveness questionnaire.

Ⅳ. DW 시스템 환경의 대응방안

4.1 비정형 데이터 분석을 위한 통합 아키텍처

앞에서 DW 시스템에 대한 문제점, 그리고 비정형 데이터 분석의 필요성에 대해서 고찰하였다. DW 시스템에서 비정형 데이터를 분석하기 위해서는 먼저 비정형 데이터를 수집하고, 적재하는 프로세스가 선행되어야 한다. 이런 프로세스를 어떻게 구성하고 H/W나 S/W적인 부분을 유기적으로 최적화된 아키텍처로 구현하는 것에 따라 비정형 데이터를 분석하기 위한 중요한 단계이다. 이전에 기존 사례나 많은 연구에서 DW 시스템에서 비정형 데이터의 수집과 적재에 관해 다양하고, 많은 제안들이 있었다. 본 연구의 핵심 포인트는 기존의 DW 시스템의 확장성과 DW 시스템에서 비정형 데이터 처리를 위한 해법에 초점을 맞춰 여러 제안들을 검토하여 최적화된 DW 시스템 고도화 방안을 제시 하고자 한다. <그림 3>은 통합 아키텍처에 대한 특징 및 효과를 정리 했다. DW, Hadoop, Data Lake를 통합하여, 문제점 및 한계를 극복하기 위한 방안으로 보완한 구성도이다.

빅데이터를 DW 시스템과 통합했을 때 가장 큰 효과는 기존 정형 데이터 분석뿐만 아니라 비정형 데이터분석 능력을 확보할 수 있다는 장점이 있다. Hadoop 플랫폼 도입으로 데이터 분산저장, 처리속도 향상, 대용량 데이터 처리 등 다양한 기능을 제공하고, DW 한계를 극복한 방안을 제시한다는 것이다. 이런 근거를 바탕으로 DW 시스템의 비정형 분석을 위한

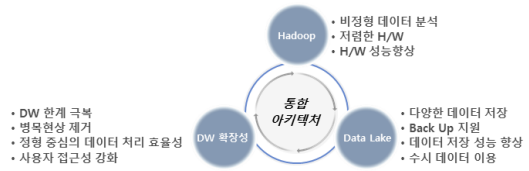

그림 3. 통합 아키텍처 특징 및 효과
Fig. 3. Unified architecture features and effects.

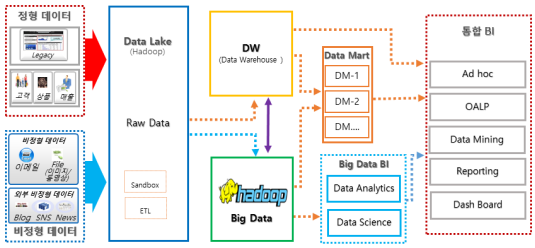

그림 4. DW 시스템에서의 비정형 데이터 분석 아키텍처
Fig. 4. Unstructured Data Analysis Architecture in DW Systems.

아키텍처 구성 방안을 DW 영역과 빅데이터 영역으로 통합하여 제시한다. <그림 4>는 원천시스템에서 DW 영역으로 데이터를 직접 수집, 적재하는 방식에서 Data Lake를 적용하여 정형, 비정형 데이터를 원시 데이터 형태로 장기적인 보관과 데이터 활용도를 극대화 시킨 아키텍처 구조이다. 또한, 비정형 데이터 분석을 위하여 빅데이터 분석을 가능하게 했다. 아키텍처 구성은 원천데이터를 수집하여 저장하는 저장 아키텍처 영역과 각 업무 특성을 감안한 DW 영역 및 빅데이터 영역으로 구분하여 각 모듈별로 그룹화 했다.

4.2 저장 아키텍처

DW 시스템에서 데이터 분석을 위하여 ODS (Operational Data Store)라는 원천 시스템에서 데이터를 수집하여 저장하는 프로세스, 즉 운영데이터를 저장하는 모듈을 구축하여 처리하고 있다. ODS는 단독으로 분석모듈로서 사용되고, DW에 데이터 소스를 전달하는 역할도 한다.

<그림 5>는 DW 기준에서의 데이터 저장 아키텍처이다. ODS에서 1차 가공 형태로 데이터를 저장하는 프로세스로 데이터의 사용에 따라 관리 주기가 유동적이다. ODS의 기본적인 운용 환경이 Data를 장기간 보관하여 분석을 위한 DW 환경이 아닌 시계열적인 주기에 따라 삭제, 생성이 반복된다는 것이다. 대용량의 데이터 처리와 정형 데이터 및 비정형 데이터를 장기적으로 보관하여 특정 요구에 따라 수시로 분석이 필요할 경우 언제든 재사용 할 수 있는 구조의 아키텍처 구성이 필요한 것이다. Data Lake의 특성이 바로 이런 부분을 해결할 수 있는 대안으로 아키텍처를 구성하였다.

Data Lake 구성은 기존 DW 시스템에서 정형 원천 데이터를 Data Mart에 적재하는 아키텍처에서 변형되어 데이터 원천을 Hadoop 플랫폼을 이용하여 저장하는 방식을 취하고 있다. Data Lake의 구성을 보면, 정형, 비정형 원천데이터를 1차로 Data Lake라는 저장소에 보관을 하는 방식이다. 기존 DW시스템에서 Tape Backup 처럼 장기 데이터 보관 의미도 있지만,

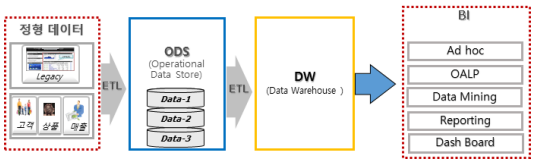

그림 5. 저장 아키텍처
Fig. 5. Storage architecture.

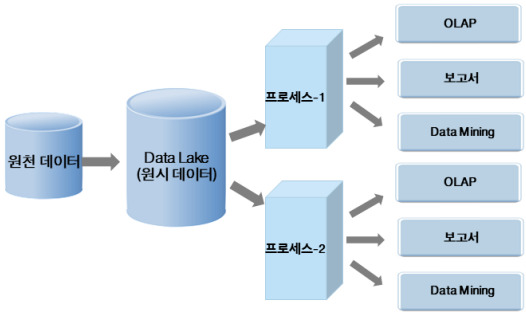

그림 6. Data Lake 아키텍처
Fig. 6. Data Lake Architecture.

Data Lake의 활용은 필요할 때 마다 꺼내서 분석할 수 있고, 데이터 저장 속도나 비정형 데이터 처리 등 많은 효과가 있다.

비정형 분석을 위한 아키텍처에서 프로세스-1, 2의 의미는 DW 시스템이 될 수 있고, 빅데이터 분석을 위한 시스템도 가능하다. 즉, Data Lake를 활용하여 대량의 데이터와 정형, 비정형 데이터 처리도 가능하다. 가장 큰 장점으로는 기존의 관계형 데이터베이스나 DW 시스템, 그리고 BI 솔루션 등과 큰 플랫폼 변경 없이 접목이 가능 하다는 것이다. 본 연구에서도 이런 특징으로 Data Lake 아키텍처를 적용했다.

4.3 DW 분석 아키텍처

DW 아키텍처 부분은 일반적인 정형 분석을 위한 시스템으로 구성했다. 기존 ODS에서 데이터를 수집하고, 분석 및 적재를 하는 형식은 기존의 시스템과 동일하지만, 본 연구에서 제안하는 구성은 Data Lake에서 Raw Data수준에서 수집하는 형태로 처리했다. 그리고 DW 시스템내부의 처리는 각 주제영역별로 구분하여 데이터를 가공하는 구조이다. 전통적인 DW 시스템 구조에서는 원천데이터 수집 경로가 Data Lake에서만 획득하여 분석을 하지만, 비정형 데이터를 분석하기 위해서는 빅데이터의 분석 자료를 정형화하여 DW 각 주제영역의 데이터와 결합하여 최종 분석 결과를 도출하는 구조로 구성했다.

DW 분석 아키텍처 구성도는 각 원천의 빅데이터 분석 자료와 Data Lake의 Raw data를 수집하여 분석하는 아키텍처이다. 해당 아키텍처의 장점은 우선 비정형 데이터를 정형화하여 DW 시스템의 한계를 극복했다. 또한, DW 시스템의 고질적인 문제점인 Batch 처리에서의 병목현상을 제거하고, 다양한 분석의 효과를 거둘 수 있다는 것이다. DW 영역에 초점을 뒀기 때문에 빅데이터의 데이터 흐름에서는 단순하게 표기

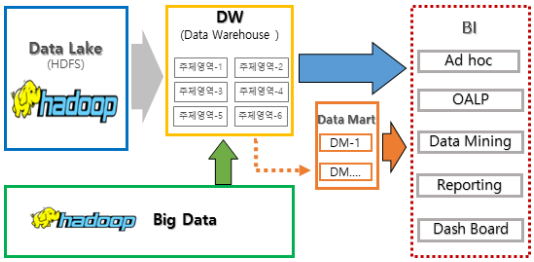

그림 7. DW 분석 아키텍처
Fig. 7. DW Analysis Architecture.

하였다. 그리고 BI 영역의 분석은 Data Mart를 통한 분석과, DW 시스템의 자료를 직접 분석하는 두 가지 패턴으로 처리된다.

4.4 빅데이터 분석 아키텍처

빅데이터 아키텍처 부분은 DW 시스템의 아키텍처와 같은 데이터 흐름으로 구성하였다. DW 시스템과 같이 원천의 데이터를 Data Lake를 통해서 수집하여 분석하는 구조로 정형 데이터는 DW 및 빅데이터에서 직접 수집한다. 빅데이터의 세부적인 Ecosystem 구성은 본 연구에서는 배제하고 DW 시스템 데이터와 빅데이터 데이터를 결합하는 방식의 아키텍처에 주안점을 두었다.

빅데이터 데이터 처리 아키텍처 구성은 DW 시스템의 정형 데이터와 빅데이터의 Raw Data인 비정형 데이터를 수집하여 분석을 위한 BI 솔루션과 결합하는 구조이다. 빅데이터 자체적인 분석결과를 Visualization하여 다양한 분석 결과를 도출해내는 아키텍처인 것이다. 실시간 분석이나 비정형 분석을 빅데이터와 DW영역에서 상호 유기적으로 결합하는 시너지 효과를 얻을 수 있다.

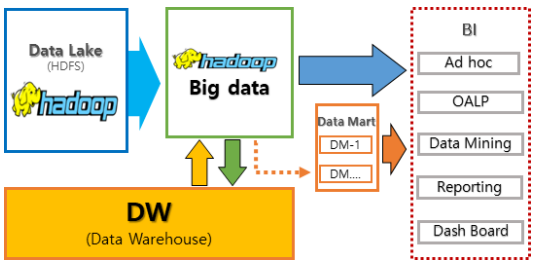

그림 8. 빅데이터 분석 아키텍처
Fig. 8. Bigdata Analysis Architecture.

V. 제안 Architecture 비교 분석

5.1 비교 분석

DW 시스템을 운용하면서 비정형 데이터 분석을 위한 다양한 변화에 대한 새로운 아키텍처 구성에 대해서 특징을 알아보기 위해 다음 사례를 분석해볼 필요가 있다. 먼저 현재 DW 시스템을 구축하여 운용하고 있는 것을 가정하고, 비정형 데이터 처리를 위해 빅데이터 시스템을 추가로 도입하여 병행으로 운용하는 시스템 구성도를 검토할 필요가 있다.

이 방안은 기존의 DW 영역과 빅데이터 영역으로 구분하여, 원천 데이터를 각 영역에 맞게 수집, 적재하여 별도의 시스템처럼 이원화하여 분석하는 방식이다. 이런 구성의 장점은 기존시스템, 즉 DW 시스템에는 전혀 영향이 없고, 신규로 빅데이터시스템을 구축하는 것이다. 최근에도 빅데이터 분석의 수요나 정책에 따라 빅데이터 부분을 분리하여 구축하고 있다. 특정의 업무프로세스는 빅데이터 영역에서 1차 가공된 형태로 DW 시스템에 원천 데이터로 제공되어 기존 DW 시스템의 병목현상이나 시스템 부하를 줄일 수 있는 점도 개선된 효과이다. 단점으로는 데이터 가공이 이원적으로 분리되어 통합의 의미보다 개별 시스템 구축 효과만 얻을 수 있는 한계가 있다는 것이다. 또한, BI 영역에서 데이터 분석이 별도의 환경에서 개별로 처리된다는 점이다. 문제는 기존 DW 시스템에 대한 확장성이 고려되지 않았고, 데이터 결합을 통한 다양한 분석에는 부족하다는 것이다. DW 시스템의 최적화된 솔루션과 빅데이터의 정제 데이터가 서로 불일치하는 경우도 있을 수 있다.

통합 유형-1과 비슷한 사례는 국내 모 기업에서 비용 절감에 대해서 이미 검증된 바 있다. 통신사의 CDR(Call Detail Record) 데이터, 즉, 통화관련 정보량이 급격히 증가하고 있는 상황에서 데이터 처리에 대한 기존 DW 시스템의 Batch 프로세스 구간에서 병목현상이 발생되고 있었다.

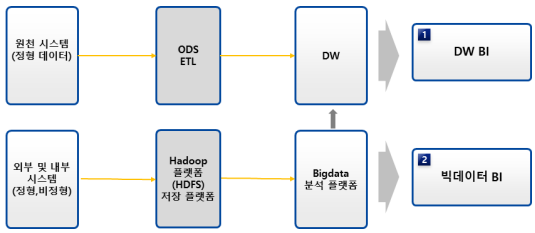

그림 9. DW 통합 유형-1
Fig. 9. DW Integration Type - 1.

표 6. DW 통합 유형-1의 특징
Table 6. Features of DW Integration Type-1.

	DW System	Hadoop
Source	- 정형 데이터	- 정형, 비정형 데이터
Storage	- DW(RDBMS)	- 분산처리(HDFS)
Data Flow	- 원천 ⇒ DW ⇒ BI - Hadoop 분석 ⇒ DW⇒BI	- 원천 ⇒ Hadoop 분석 ⇒ BI - 원천 ⇒ Hadoop 분석 ⇒ DW
Processing Speed	- Batch 처리 감소 - 병목현상 해소	- 분산 처리
Data Storage	- Tape Backup	- 분산 Disk
Characteristic	- 정형분석의 최적화된 DW고유 업무 수행 - 기존DW한계 극복(비정형, 대량 데이터 분석) - 비용 효율적 - 시스템의 이원화 - 데이터 품질확보 등 거버넌스 문제	

이에 대한 해결방안과 데이터 증가에 따른 시스템 증설비용 등을 절감하기 위해 Hadoop 기반의 x86 서버로 통화호 데이터를 기존 DW 시스템에서 분리하여 신규로 빅데이터 플랫폼을 적용하였다. 이를 통해 해당 기업은 가입자 분석 시스템의 성능개선과 RDBMS 기반의 아키텍처와 비교했을 때 TCO(Total Cost of Ownership) 기준으로 상당한 비용 절감 효과를 얻었다. 또한, 성능 개선에서 ETL 병목현상 해소 등의 많은 부분에서 개선하였다. 이에 반해 통합 유형-1의 다른 측면은 시스템의 분리로 인한 데이터 거버넌스 체계 확보에서 문제점도 발견되었다. 즉, 데이터의 신뢰성과 품질에 관한 문제에서 DW 시스템 데이터와 빅데이터에서 분석한 데이터와의 정합성에 관한 사항이다. 이로 인한 데이터 보정작업이 수반되고, OLAP 분석을 별도 각각의 시스템에서 수행하는 경우도 있다. 또한, 특정의 데이터 분석이 필요할 경우 데이터 수집과 적재부터 처음부터 재작업 하는 경우가 다수 존재한다. 다음은 통합 유형-1의 문제점을 해결하고 좀더 DW 시스템을 확장할 수 있는 방안을 토대로 두 번째 제안 내용을 검토하려고 한다. 이번 제시안의 중요한 부분은 데이터 저장 방식을 DW 시스템과 빅데이터의 Hadoop 플랫폼을 단일 시스템으로 통합하여 처리하는 프로세스로 유형-1과 다르게 원천데이터를 하나의 시스템에서 수집, 적재하고 분석하는 시스템이다.

DW 통합 유형-2의 가장 큰 특징은 저장 플랫폼을 Hadoop 기반의 분산처리 시스템을 이용하여 통합 저장소를 구축하여 대용량의 데이터와 정형, 비정형 데이터를 빠른 시간에 처리하는 개선 효과를 위한 방안이다.

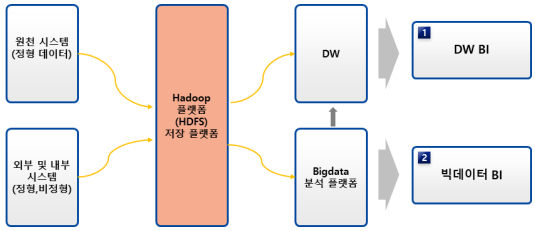

그림 10. DW 통합 유형-2
Fig. 10. DW Integration Type - 2.

통합 유형-2의 특징은 데이터 저장프로세스의 변화로 유형-1보다 데이터 품질이나 이중화의 단점은 해소되었다. 하지만 아직도 최종적인 DW 시스템에서 비정형 분석의 단일화는 실현하지 못하는 시스템 구성안이다. 여기에 Hadoop 기반의 데이터 저장 방식이 Data Lake 솔루션 지원 이전의 방식으로 구성되어 있다. 물론 기존 Hadoop 장비를 운용하고 있는 기관에서는 이런 방식 또한 나쁘지 않다. 하지만, DW 통합 유형-1과 통합유형-2를 기준으로 데이터 거버넌스를 확보하고, 대용량의 데이터 처리를 위한 비용 절감, 그리고 비정형 데이터를 DW시스템에서 통합하여 구축하는 기본 구성이다. 이런 변화의 주된 요인을 분석해보면 다음과 같은 특징으로 분류할 수 있다.

첫째, 기존 운영되고 있는 DW 시스템을 적극 활용하여 안정적으로 의사결정지원에 반영하고 있다는 점이다. 둘째, DW 시스템의 문제점인 비정형의 데이터 처리를 좀더 유연하게 분석할 수 있는 새로운 방안이

표 7. DW 통합 유형-2의 특징
Table 7. Features of DW Integration Type-2.

	DW System	Hadoop
Source	- 정형, 비정형 데이터	
Storage	- 분산처리(HDFS)	
Data Flow	- Hadoop저장 ⇒ DW ⇒ BI - Hadoop분석 ⇒ DW ⇒ BI	- Hadoop저장 ⇒ Hadoop ⇒ BI - Hadoop저장 ⇒ Hadoop분석 ⇒ DW
Processing Speed	- Batch 처리 감소 - 병목현상 해소 - 분산 처리	
Data Storage	- 분산 Disk	
Characteristic	- 유형-1의 장점 - 저장 플랫폼 단일화 - 데이터 품질확보 등 거버넌스 문제 해소	

도입된 점이다. 그리고 이런 두가지의 방식을 결합한 하이브리드 방식의 데이터 분석이 가능하게 했다는 점이 주된 특징으로 보여진다. 이에대한 고려사항으로는 인프라 측면에서 확장용이성, 비용효율성, 자원 할당 및 관리의 용이성이 중요한 핵심사항이다.

Ⅵ. 결 론

최근들어 빅데이터 시장이 도입 초기 단계를 넘어 전반적인 산업분야에 적극 활용되고 있는 시점에서, 과거 DW 시스템의 안정적인 운영과 결과에 어느정도 만족을 하고 있던 많은 기업들이 빅데이터 분석을 위한 새로운 시스템을 적용하는 과정에서 가장 합리적이고, 효율적인 시스템 운용을 위해 많은 고민을 하고 있다. 대량의 데이터를 수년간 축적해온 기업에서 시기별 추이분석을 위한 방대한 데이터 분석을 위해서는 기존의 DW 시스템의 마트에서 데이터를 수집, 가공하는 방안이 더 효율적이고, 시간이나 비용적인 측면에서 유리하다는 것이다. 즉, 기존 운영계의 데이터를 정보계로 매일 적정한 분석을 위해 1차 집계를 하는 DW 시스템의 정보데이터가 더욱 값진 결과로 추출할 수 있다는 것이다. 반대로 대량의 다양한 데이터를 빅데이터 시스템에서 분석한 결과를 특정의 1회성 데이터로 사용하고 폐기하는 것 보다 연관 분석을 위한 또 다른 데이터 활용을 위해 적절하게 분배하여 DW 시스템으로 정보제공을 하는 것이 향후 운영적인 측면이나, 비용적인 면에서 효율성을 제공한다는 것이다.

이런 변화에 대해서 이전에 DW 시스템을 연계하여 통한 데이터 분석을 하는 기업들이 점차 증가하는 추세이고, 어떤 경우는 기존의 DW 시스템을 재구축 또는 고도화 추세의 변화도 있다. DW 시스템을 개선하기 위한 목표는 다음과 같은 기준에 부합되어야 새로운 변화에 적응 할 수 있다.

첫째, 기본적으로 DW 시스템의 고도화 및 확장성을 개선 적용하였다.

둘째, 빅데이터 부분을 수용 할 수 있어야 한다.

셋째, 비용 절감 및 프로세스 성능개선이 보장 되어야 한다.

넷째, BI 솔루션의 적용이 원활해야 한다.

즉, 원천 데이터의 재활용성을 확대하고, 정형, 비정형 다양한 데이터를 저장할 수 있는 분산처리 플랫폼을 구축과 기존 DW 시스템의 범용성에 더욱 초점을 두고 고도화 작업을 진행해야 한다. 각 기업체에서 매년 증가하는 데이터처리와 사회 관계망 서비스의 발전으로 다양한 채널에서 유입되는 정보의 분석이 절실히 요구되는 시대가 되었다. 이를 해결하기 위해 신규로 빅데이터 시스템을 구축하는 경우도 많지만, 기존 DW 시스템을 확장하여 Hadoop 기반의 대용량 데이터 처리와 유지보수 비용의 절감 효과, 그리고 정형, 비정형 데이터를 상호 유기적으로 분석할 수 있는 통합 시스템 구축이 가져오는 시너지 효과는 상당히 크다. 기존 DW 시스템과 통합으로 DW 성능 향상과 DW 비용의 절감 효과를 가져 왔으며, 실시간 고객 분석을 처리할 수 있는 환경을 구축할 수 있다는 장점으로 이제, 빅데이터 시대에서 DW 시스템을 바라보는 시각이 새로운 대체가 아닌 공존의 시스템으로 운용하는 것도 검토해 볼 시기이다.

References

[1] D. Jang, *Technology that works with big data*, Hanbit Media, 2014.

[2] Y.-W. Jeong and S.-K. Cheong, "Data warehouse (DW) appliance technology trend analysis," *Korea Inst. Inf. Technol. Mag.*, 2012.

[3] McKinsey, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey & Company, 2011.

[4] Dbguide, *Big Data Age BI is the answer to the advanced DW*, encore, 2013.

[5] S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," *J. Busin. Res.*, vol. 69, no. 2, pp. 897-904, Feb. 2016.

[6] W. H. Inmon, *Building the Data Warehouse*, Hongrung Publishing, 1997.

[7] Philip Russom, *Integration of Hadoop with BI / data warehousing*, TDWI Reseach, 2013.

[8] J. D. Warren, K. C. Moffitt, and P. Byrnes, "How Big Data will change accounting," *Accounting horizons*, vol. 29, no. 2, pp. 397-407, Jun. 2015.

[9] B. Cha, S. Park, B.-C. Shin, and J. Kim, "A pilot study on bigdata-based data lake concept for business intelligence," in *Proc. KICS Winter Conf.*, pp. 1300-1301, Jan. 2018.

김 재 형 (Chae-Hyeong Kim)

2018년 2월 : 숭실대학교 정보과학대학원 IT경영학 전공(공학석사)
2018년~현재 : 숭실대학교 IT정책경영학과 박사과정
<관심분야> Bigdata, AI, 스마트 시티, IOT, DW

[ORCID:0000-0002-7853-4147]

장 동 원 (Dong-Won Jang)

2006년 2월 : 숭실대학교 정보과학대학원 소프트웨어공학 전공(공학석사)
2017년~현재 : 숭실대학교 IT정책경영학과 박사과정
<관심분야> 스마트시티, 빅데이터, 지능형영상감지, 영상관제솔루션

[ORCID:0000-0002-9639-4619]

김 석 (Seog Kim)

2013년 8월 : 숭실대학교 정보과학대학원 IT경영학 전공(공학석사)
2018년~현재 : 숭실대학교 IT정책경영학과 박사과정
<관심분야> 블록체인, 비즈니스모델, 미디어서비스

[ORCID:0000-0002-5343-3411]