한국빅데이터학회지
제1권 제2호, 2016, pp. 91-102

전통적 환경과 빅데이터 환경의 데이터 자원 관리 비교 연구

A Study on Data Resource Management Comparing Big Data Environments with Traditional Environments

박주석[1†] · 김인현[2]

경희대학교 경영대학[1]
투이컨설팅[2]

요 약

전통적인 환경에서 데이터 생명주기는 데이터-정보-지식-지혜 전환과정으로 요약된다. 반면에 빅데이터 환경에서 데이터 생명주기는 데이터-통찰-실행 전환과정으로 요약된다. 이러한 전환과정의 차이점은 데이터 생명주기를 지원하는 데이터 자원 관리에도 변화를 요구한다. 본 논문에서는 전통적인 데이터 자원 관리와 비교하여 빅데이터 환경을 위한 데이터 자원 관리를 연구한다. 특히 빅데이터 자원관리를 위한 주요 구성요소를 제안한다.

■ 중심어 : 빅데이터, 데이터 자원 관리, 데이터 생명주기, 데이터 레이크

Abstract

In traditional environments we have called the data life cycle DIKW, which represents data-information-knowledge-wisdom. In big data environments, on the other hand, we call it DIA, which represents data-insight-action. The difference between the two data life cycles results in new architecture of data resource management. In this paper, we study data resource management architecture for big data environments. Especially main components of the architecture are proposed in this paper.

■ Keyword : Big Data, Data Resource Management, Data Life Cycle, Data Lake

Ⅰ. 서 론

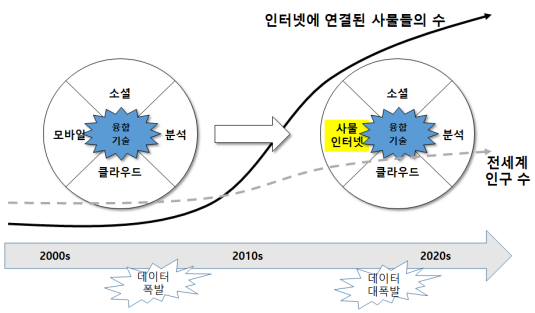
정보자원은 하드웨어 자원, 소프트웨어 자원, 데이터 자원 등으로 분류될 수 있다. 초기에는 하드웨어 자원이 워낙 비싸기 때문에 하드웨어 자원의 효율적 활용에 많은 관심을 가졌다. 시간이 지나면서 실제로 업무를 자동화시키는 소프트웨어 자원에 집중하기 시작하였다. 기업에 적용된 소프트웨어 자원에 따라 기업 생산성에 확실한 차이가 생겼으며, 모든 기업들이 소프트웨어 자원에 집중하였다. 또한 잘못된 소프트웨어 자원은 엄청난 기업 리스크를 초래할 수도 있다는

2016년 8월 25일 접수; 2016년 8월 31일 게재 확정.
† 교신저자 jspark@khu.ac.kr

점을 인식하기도 했다. 이제는 거의 모든 업무가 정보화됨에 따라 정보시스템을 통해 생성되는 데이터 자원에 관심을 갖기 시작했다. 데이터 자원을 연계하고 통합하면서 새로운 가치를 창출할 수 있게 됨을 확인하게 되었고 데이터 자원을 분석하여 빠르고 정확한 의사결정을 가능하게 되었다. 이런 관점에서 알리바바 회사의 마윈 회장은 데이터기술(DT: Data Technology)을 강조하고 있다.

최근에 빅데이터 환경이 도래하면서 데이터 자원에 더욱 관심을 갖게 되었다. 소위 스맥(SMAC: Social, Mobile, Analytic, Cloud의 약자) 환경이 도래하면서 데이터는 폭발적으로 증가하고 있다. 기업들은 빅데이터 관리를 비용으로 보지 않고 기회로 보기 시작했다. 예를 들어 기업들은 과거에 고객 거래 데이터 등의 구조적 데이터만을 관리하고 분석하였으나, 최근에는 고객 접촉데이터 등의 비구조적 데이터도 관리하고 분석하기 시작하였다. 이러한 빅데이터 분석을 통해서 경영 성과를 획기적으로 높였던 여러 가지 사례가 제시되었다.

미래 정보기술 환경을 예측해보면 데이터 자원 관리는 더욱 중요해진다. 시스코(CISCO) 회사의 챔버스 회장은 '2020년에는 25억 명의 사람과 370억 개의 사물이 인터넷에 연결되고 2030년에는 500억 개의 사물이 인터넷에 연결되어 혜택을 줄 것이다.'라고 예측하였다. 사물인터넷과 빅데이터는 미래를 향한 중요한 2가지 축이 될 것이다. 사물인터넷은 앞에서 센서 등을 비롯한 다양한 디바이스를 통해서 실시간으로 엄청난 데이터를 생성해 낼 것이고 빅데이터는 뒤에서 생성된 데이터를 가공하고 분석하여 사물인터넷이 지능화될 수 있도록 지원할 것이다. 결국 사물인터넷과 빅데이터의 두 축이 시너지가 될 수 있도록 새로운 관점의 데이터 자원 관리가 필요하다.



〈그림 1〉 스맥, 사물인터넷(IoT) 그리고 데이터 대폭발

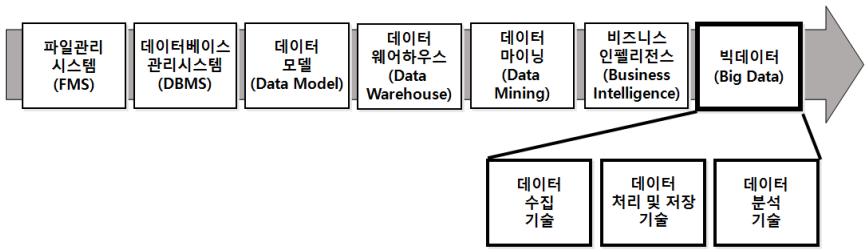
본 논문에서는 빅데이터 환경하에서 데이터 자원 관리에 대해서 연구하고자 한다. 먼저 전통적인 데이터 생명주기와 빅데이터 환경하의 데이터 생명주기를 비교하고, 이를 근거로 새로운 데이터 자원 관리를 제안하고자 한다. 특히 빅데이터 자원관리를 위한 주요 구성요소를 제안한다.

Ⅱ. 데이터의 이론적 고찰

2.1 데이터 기술의 역사 그리고 빅데이터 기술

지난 50년간 데이터의 역사를 살펴보면, 주요 거래데이터를 저장하기 위한 파일관리시스템(FMS: File Management Systems), 다양한 거래데이터를 통합적으로 관리하기 위한 데이터관리시스템(DBMS: Database Management Systems), 그리고 이러한 데이터의 중복성을 없애고 일관성을 유지하기 위한 데이터 모델(Data Model), 업무처리 관점이 아닌 사용자 관점으로 분석하기 위한 데이터 웨어하우스(Data Warehouse), 수많은 데이터에서 가치있는 정보를 찾아내기 위한 데이터 마이닝(Data Mining), 기업 지능 수준을 높이기 위한 정보화체계인 비즈니스 인텔리전스(BI: Business Intelligence) 등의 발전이 있었다. 이러한 데이터 기술의 발전은 주로 기업 데이터이면서 구조화된(structured) 데이터에 집중되었다.

오히려 1990년대에 들어오면서 BPR, ERP, 6시그마 등의 경영혁신 기법이 일반화되면서 데이터 기술보다는 프로세스 기술에 대한 연구가 활성화되었다. 시스템개발 방법론도 데이터 중



〈그림 2〉 데이터 기술의 역사 그리고 빅데이터 기술 분류

심보다는 프로세스 중심의 방법론이 활성화 되었다. 실질적으로도 구조화된 데이터 기술 수준은 어느 정도 안정이 되었기 때문에 2010년대에 들어오면서 데이터 아키텍처와 데이터 품질에 대한 연구만이 활성화되었다.

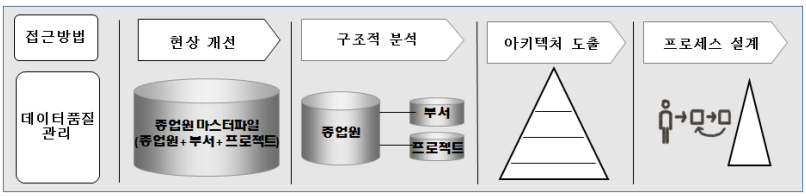
빅데이터에 관심을 갖기 시작한 것은 2000년대 중반부터이다. 일반 기업과 달리 아마존과 구글은 인터넷상에서 엄청나게 많은 데이터를 관리해야 하기 때문에 새로운 기술이 필요했고 그 당시 대표적인 기술이 구글의 '빅테이블'과 아마존의 '다이나모'이었다. 이러한 기술들은 오픈 환경의 분산처리시스템으로 일반화되었는데 하둡이 대표적이다. 하둡을 대규모 데이터 처리에 활용한 대표적인 사례로는 페이스북이 있다. 페이스북은 하둡을 비롯한 오픈소스(Open Source)를 적용하여 데이터센터를 설립하였다. SNS 데이터 분석은 하둡 위에 하이브(Hive)로 구성된 데이터 웨어하우스를 사용하였다. 초기에는 빅데이터를 어떻게 저장하고 관리할 것인지가 중요한 이슈이었으나, 시간이 지날수록 빅데이터를 어떻게 분석하고 활용할 것인지가 더 중요한 이슈가 되었다. 일반적으로 빅데이터 기술은 수집 기술, 처리 및 저장 기술, 그리고 분석 기술로 구분된다.

2.2 데이터 품질관리의 역사

데이터 품질관리의 역사를 살펴보면, 사용자(User) 관점에서 시작하여 모델러(Modeler) 관점으로, 아키텍트(Architect) 관점으로, 그리고 최근에는 거버너(Governor) 관점으로 발전되었다. 처음에는 사용자가 사용하는 데이터의 품질을 높이기 위해서 데이터 값을 주로 분석하였다. 데이터 값을 개선해도 결국 데이터 구조를 개선해야 근본적인 개선이 되었기 때문에, 나중에는 데이터 값보다 데이터 모델링에 많은 노력을 기울였다. 데이터 모델링을 통해서 각 업무영역에서 데이터 품질이 높여졌지만, 전사 차원의 데이터 품질을 위해서는 데이터 아키텍처 관리가 중요해졌다. 전사차원의 데이터 품질을 지속적으로 유지하기 위해서는 데이터 아키텍처와 함께 데이터 거버넌스가 필요해졌다. 최근에 빅데이터 환경이 도래함에 따라, 구조적 데이터뿐만 아니라 비구조적 데이터를 위한 품질관리도 중요해졌다.

2.3 정보자원관리와 데이터 아키텍처

정보자원관리(Information Resource Manage-



〈그림 3〉 데이터 품질관리의 역사

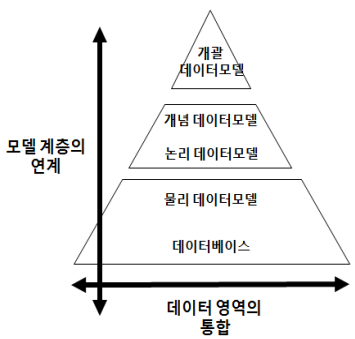
	업무 아키텍처	응용 아키텍처	데이터 아키텍처	기술 아키텍처	보안 아키텍처
경영자	BV1 조직구성도/정의서 BV2 업무구성도/정의서	AV1 응용서비스구성도 /정의서	DV1 데이터구성도 /정의서	TV1 기반구조구성도 /정의서	SV1 보안정책 SV2 보안구성도 /정의서
책임자	BV3 업무기능관계도 /기술서 BV4 업무기능분할도 /기술서	AV2 응용서비스관계도 /기술서 AV3 응용기능분할도 /기술서	DV2 개념데이터관계도 /기술서 DV3 데이터교환기술서	TV2 기반구조관계도 /기술서	SV3 보안관계도 /기술서
설계자	BV5 업무절차설계도 /설계서	AV4 응용기능설계도 /설계서	DV4 논리데이터모델 DV5 데이터교환설계서	TV3 기반구조설계도 /설계서 TV4 시스템성능설계서	SV4 관리보안설계서 SV5 물리보안설계서 SV 6기술보안설계서
개발자	BV6 업무매뉴얼	AV5 응용프로그램목록	DV6 물리데이터모델	TV5제품목록	SV7 보안매뉴얼

〈그림 4〉 우리나라 정부의 엔터프라이즈 아키텍처 프레임워크

ment) 개념은 1980년대부터 시작되었으나, 2000년대에 들어오면서 많은 관심을 갖기 시작하였다. 그 이유는 기업이나 기관이 정보화가 많이 될수록 정보시스템이 복잡해질수록 정보자원관리가 중요해지기 때문이다. 정보자원관리를 지원하기 위해 엔터프라이즈 아키텍처(EA: Enterprise Architecture)라는 기법을 많이 적용하고 있다. 엔터프라이즈 아키텍처는 업무아키텍처, 응용아키텍처, 데이터 아키텍처, 기술아키텍처로 구성되어 있다. 예를 들면 우리나라 정부는 5개의 뷰(view)로 구성된 공공 엔터프라이즈 아키텍처를 법제화하여 적용하고 있다.

많은 전문가들이 4가지 아키텍처 중에서 데이터 아키텍처가 가장 중요하다고 한다. 첫 번째 이유는 업무프로세스, 애플리케이션 프로그램, 정보시스템 기술 등은 지속적으로 변하지만, 데이터 구조는 잘 변하지 않기 때문이다. 안정적인 데이터구조로 구성되어 있는 정보시스템은 아무리 기업 환경이 바뀌어도 안정적인 형태를 유지할 수 있다. 두 번째 이유는 타 아키텍처에 비해서 관리하기가 복잡하지 않다. 즉, 데이터와 관련되어 있는 애플리케이션 프로그램 구조는 매우 복잡하지만, 데이터 구조는 간단하다. 세 번째 이유는 조직이 경쟁력을 갖추기 위해서는 비즈니스 분석 능력이 높아져야 한다. 비즈니스 분석을 위해서는 데이터 자원의 가시정이 가장 중요하다.

데이터 아키텍처는 최상위의 개괄적인 수준에서부터 데이터베이스 수준까지 데이터에 관한 모든 구조를 통합하여 연계하고, 업무 및 기술들 타 아키텍처와의 전체적인 관계를 정립한다. 즉, 기업의 전사 데이터 모델을 5개의 계층으로 구분하여, 개괄 데이터 모델, 개념 데이터 모델, 논리 데이터 모델, 물리 데이터 모델, 데이터베이스로 보여준다.

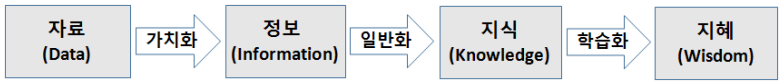


〈그림 5〉 일반적인 데이터 아키텍처 프레임워크

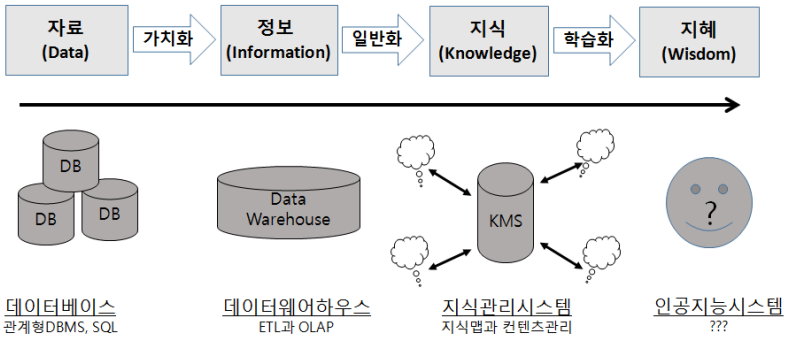
Ⅲ. 데이터 생태계의 변화

3.1 전통적인 환경하의 데이터 생명주기

전통적인 데이터 생명주기는 DIKW(Data, Information, Knowledge, Wisdom)이다. 데이터를 가공하면 정보가 되고, 정보가 축적되면 지식이 되며, 지식이 일반화되면 지혜가 된다는 것이다. 지혜를 갖추면 최선의 행동을 할 수 있다. 데이



〈그림 6〉 전통적인 데이터 생명주기



〈그림 7〉 전통적인 데이터 생명주기 구현방안들

터는 최종적으로 의도한 행동을 이끌어낼 때 목적을 달성하게 된다.

기업 또는 기관이 데이터를 보관하고 활용하는 것은 구성원이 조직의 목적에 맞는 행동을 하도록 유도하기 위해서이다. 이를 위하여 기업은 데이터 웨어하우스(DW: Data Warehouse)와 지식관리시스템(KMS: Knowledge Management System)을 구축하여 운영하여 왔다. 데이터 웨어하우스는 거래 데이터를 통합하여 보관하고 목적에 맞게 분류하여 제공하는 역할을 수행한다. 지식관리시스템은 거래 데이터로 기록할 수 없는 경험이나 사례들을 보관하고 공유할 수 있도록 한다.

데이터 웨어하우스와 지식관리시스템에 많은 투자를 했지만 그다지 성공적이지 못했다. 많은 데이터와 지식을 제공했음에도 구성원들이 조직의 목표에 부합하는 행동을 하게 하는 데는 그다지 성공적이지 않았다. 영업 성공사례를 지식으로 등록하고 쉽게 찾아볼 수 있게 했지만 잘 활용되지 않았다. 고객의 거래 데이터를 분석해서 의미있는 정보를 제공해도 현장의 영업직원은 자신의 경험이나 지식에 의존하여 행동하고 있다.

전통적 데이터 생명주기가 원활하게 작동하지 않는 원인은 다음과 같다. 첫째, 데이터에서 행동으로 이어지는 시간이 너무 길다. 데이터 발생 시점부터 데이터 웨어하우스에 정보로 제공되기 까지는 통상 한 달 정도 걸린다. 사람이 정보를 확인하고 해석하여 자신의 지혜로 바뀌는 것은 사람에 따라 다르겠지만 상당 기간의 학습해야 한다. 따라서 데이터 발생 시점의 상황은 실제 행동할 시점의 상황과는 전혀 달라질 가능성이 크다.

둘째, 자료에서 정보로 가공된 이후에 지식으로 넘어가는 과정이 단절되어 있다. 정보가 제공하는 의미를 모두 파악하는 것은 불가능하다. 정보를 제공한 후에 유용한 지식들을 잃어버릴 수 있다. 또한 이러한 변환 과정은 사람의 개인 능력에 크게 의존한다.

셋째, 개인과 기업의 이해관계가 다르기 때문에 높은 지혜를 가진 사람이 기업에 유리한 행동을 반드시 하는 것은 아니다. 사람은 자신에게 유리한 행동을 선호하는 경향이 있다. 부서 수준에서도 각 부서의 최적이 전체 회사의 최적으로 결과되는 것은 아니다. 회사에 유리한 행동을 이끌어내기 위해서는 지혜를 높여주는 것

만으로는 충분하지 않다.

3.2 빅데이터 환경 하의 데이터 생명주기

빅데이터 시대가 되면서 데이터 능력이 크게 향상되었다. 첫째, 모든 상황이 데이터로 기록된다. 전통적 데이터 환경에서는 거래의 결과만 데이터베이스에 기록되었다. 빅데이터 시대에는 사람의 위치와 생각, 의견 등이 기록된다. 기계의 상태는 센서에 의해 감지되어 기록된다. 온도, 습도, 강우량 등 날씨도 순간순간 기록되고 있다. 전통적 데이터 환경에서는 현상의 10% 정도가 기록되었다면, 빅데이터는 나머지 90%의 대부분이 기록되고 있다. 따라서, 빅데이터 시대에는 하나의 사건에 대해서 더 많이, 더 자주, 더 빨리 알 수 있다.

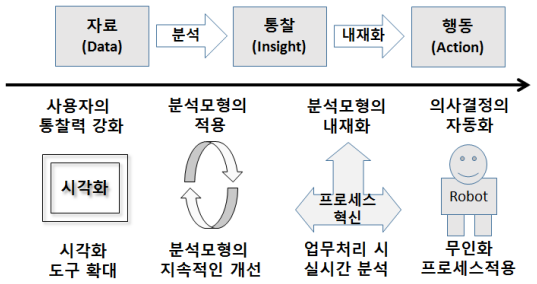
둘째, 대규모 데이터를 처리할 수 있다. 전통적 데이터 환경에서 데이터는 처리할 수 있는 규모가 제한적이었다. 또한, 발생한 데이터를 활용할 수 있으려면 수 일 이상의 기간이 필요하였다. 빅데이터 시대에는 하드웨어와 관련 소프트웨어의 비약적 발전으로 아무리 큰 데이터라도 보관하고, 분류하고, 검색할 수 있다. 또한, 데이터 발생을 바로 감지하여 실시간으로 처리할 수도 있게 되었다.

셋째, 외부 데이터를 활용할 수 있다. 전통적 데이터 환경에서는 기업이 자체 보유하고 있는 데이터만을 가공하고 분석할 수 있었다. 빅데이터 시대에는 정부와 공공기관이 보유하고 있는 공공데이터가 개방되어 자유롭게 접근할 수 있다. 또한 데이터를 수집하여 제공하는 외부의 데이터사업자가 늘어남에 따라 다양한 외부 데이터를 활용할 수 있다. 사람들이 페이스북, 트위터, 블로그 등을 보다 많이 이용함에 따라서, 일반 대중이 어떤 생각을 하고 있는가를 소셜데이터 분석을 통해 확보할 수 있다.

데이터 능력이 크게 향상됨에 따라서 데이터 생명주기는 DIA(Data, Insight, Action)로 바뀌고 있다. 기업 내부 및 외부의 데이터를 연계하여 분석할 수 있게 됨에 따라서, 데이터로부터 바로 통찰(insight)을 얻는다. 분석을 프로세스에 내재화하면, 통찰에 따라 필요한 행동을 자동으로 실행할 수 있다. 전통적은 분석은 사람이 수행하는 방식이지만, 빅데이터 환경에서는 분석 모듈을 미리 만들어서 데이터가 도착하면 자동으로 분석 모듈이 실행되고 분석 결과 값에 따라 어떤 행동을 실행할 것인가가 결정된다. 분석 결과 값에 따라 필요한 행동이 실행됨으로써, 데이터로부터 행동까지 즉시 처리되는 프로세스가 구현될 수 있게 되었다.

〈그림 8〉 빅데이터 환경하의 데이터 생명주기

빅데이터 환경의 데이터 생명주기를 구현하는 대표적인 방안은 분석모형 개발과 분석모형 내재화이다. 분석모델 개발은 분석 주제와 확보한 데이터 특성을 고려하여 적합한 분석기법을 선정한다. 분석기법 적용 타당성에 따라 선정되었던 분석 모델을 보완하거나 새로운 분석모델로 대체한다. 분석 모델의 성능은 한 번의 시도로 목표 수준을 달성하기는 쉽지 않다. 실무에 적용하고 반복적으로 개선함으로써 분석모델의 성능은 진화한다.



〈그림 9〉 빅데이터 생명주기 구현방안

분석모형 내재화는 분석모형을 업무프로세스

에 반영하는 것이다. 과거 프로세스 혁신이 기존 프로세스를 없애고 새로운 프로세스를 도출하는 것이었다면, 최근 프로세스 혁신은 기존 프로세스에 분석모형을 내재화시켜서 지능화하는 것이다. 즉, 분석 모델 결과 값에 따라서 수행해야 할 행동이 바로 이루어질 수 있도록 의사결정을 실시간으로 지원하는 것이다.

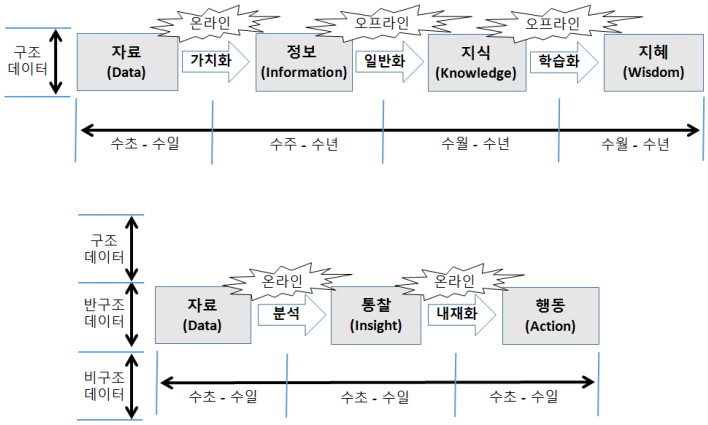
3.3 데이터 생명주기 비교 분석

전통적인 환경의 데이터 생명주기에는 온라인프로세스와 오프라인프로세스의 연계로 이루어진다. 예를 들면 데이터 생성에 사람이 직접 데이터를 입력하기도 하고 데이터 웨어하우스를 활용해서 사람이 여러 가지 분석을 별도로 하고, 분석된 내용을 근거로 새로운 지식을 지식관리시스템에 추가하기도 한다. 그리고 이러한 정보와 지식을 활용하여 의사결정이 이루어지면 사람이 직접 액션을 수행한다.

반면에 빅데이터 환경하의 데이터 생명주기는 대부분 온라인프로세스로 이루어진다. 예를 들면 센서를 통해 직접 데이터가 전달되거나 웹 크롤링을 통해 데이터를 자동 수집된다. 수집된 데이터는 자동적으로 분석되어 통찰을 주며, 많은 경우에 직접 액션을 취한다. 예를 들어 알파고의 바둑 대국을 보면, 상대방 데이터를 자동으로 수집하고, 수집된 데이터와 축적된 데이터를 분석하여 자동적으로 의사결정을 내린다.

따라서 이러한 두 가지 데이터 생명주기를 여러 가지 측면에서 비교해 볼 수 있다. 시간적인 측면에서 보면 전체 주기시간이 수주에서 수개월이었다면 수초에서 수시간으로 변화된다. 데이터 원천 차원을 보면 구조적인 데이터만을 다루었다면, 반구조적인 데이터, 비구조적인 데이터를 다룬다. 분석 모형 측면에서 보면, 통계분석이 주로 활용되었다면, 마이닝, 시각화, 최적화, 머신러닝 기법이 많이 적용되고 있다.

빅데이터 환경의 분석은 전통적 환경 분석과 비교하여 몇 가지 특징을 갖는다. 첫째, 관련성(relevance)이 대폭 향상된다. 관련성은 원인이 되는 사건과 결과로서 나타나는 상황이 얼마나 밀접하게 연관되는가 이다. 예를 들어서 특정 고객에게 어떤 상품을 권유할 때, 고객이 원하는 상품을 권유했다면 관련성이 큰 것이다. 마케팅 캠페인에서는 관련성이 높으면 전환률(conversion rate)이 높아진다. 기업 경영 활동에서 관련성을 높일 수 있다면 보다 나은 성과를 얻을 수 있다. 빅데이터는 분석을 위해 보다 광범위한 데이터를 활용할 수 있기 때문에 높은 관련성을 확보할 수 있다. 과거에는 고객의 거



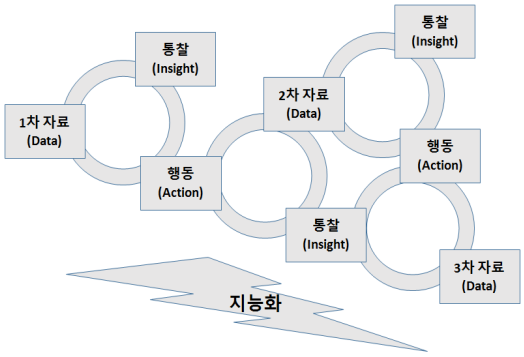
〈그림 10〉 데이터 생명주기 비교

래 데이터만으로 권유할 상품을 찾아냈다면, 빅데이터 분석 환경에서는 고객의 위치와 상태, 의견 등을 종합하여 분석할 수 있기 때문에 결과의 관련성은 상당히 높아질 수 있다.

둘째, 기민성(agility)이 높아진다. 전통적 분석은 데이터 처리 기술의 한계로 인하여 발생된 원천데이터를 데이터 웨어하우스로 모으고 다시 분석을 위한 데이터 마트로 만드는 과정에 1주에서 한 달까지 기간이 소요되었다. 따라서, 데이터를 입수하여 행동에 옮기기까지 시간이 너무 많이 걸려서 상황 변화에 즉시 대응할 수가 없었다. 빅데이터 환경에서는 다량의 데이터를 하둡 등의 병렬처리 기술을 이용하여 보다 빠르게 통합할 수 있다. 또한, 메인메모리 DBMS 기술을 이용함으로써, 생각의 속도(speed of thought)로 탐색할 수 있게 되었다. 이벤트처리 기술을 적용한다면, 특정 사건이 발생하자마자 미리 만들어진 분석 모듈이 작동하여 결과를 생성하고, 결과 값에 따라 프로세스가 즉시 수행될 수도 있다.

셋째, 분석 능력은 전략적 무기(strategic weapon)가 된다. 전통적 분석 환경에서 기업의 분석 능력은 주로 사람의 역량에 달려 있었다. 데이터의 한계와 분석 결과의 관련성과 기민성이 낮은 수준이었기 때문에 업무를 수행하는 담당자의 경험과 직관에 의하여 의사결정을 내릴 수밖에 없었다. 빅데이터 분석 환경에서는 데이터에 의한 분석 결과에 따라서 경영 의사결정을 내리고 적용할 수 있게 되었다. 따라서, 동종 기업이라면 보다 우수한 분석 모듈을 활용하는 기업이 그렇지 않은 기업보다 뛰어난 경영성과를 낼 수 있다. 그 결과로, 기업의 분석 역량은 사람이 아니라 데이터 기반 의사결정 체계의 효과성에 따라 좌우된다. 기업 경쟁력의 원천은 바로 데이터 분석 역량이 되는 셈이다.

빅데이터 시대가 도래하면서 여러 특정 비즈니스 영역에서 데이터 분석과 활용이 이루어졌다. 이러한 특정 비즈니스 영역의 분석과 활용은 시간이 지나면서 폭과 깊이가 점점 확대되고 있다. 더 나아가서 특정 비즈니스 영역의 분석과 활용에서 도출된 결과 데이터가 다른 비즈니스 영역의 데이터와 융합되면서 새로운 관점의 분석과 활용이 이루어진다. 예를 들어 공공데이터가 개방되면서 공공데이터를 가공하여 서비스하거나 공공데이터와 민간데이터를 융합하여 새로운 비즈니스 모델을 만들어내고 있다. 궁극적으로 전통적인 데이터 생태계가 서로 융합되면서 새로운 데이터 생태계를 창조하게 될 것이다.



〈그림 11〉 새로운 데이터 생태계 도래

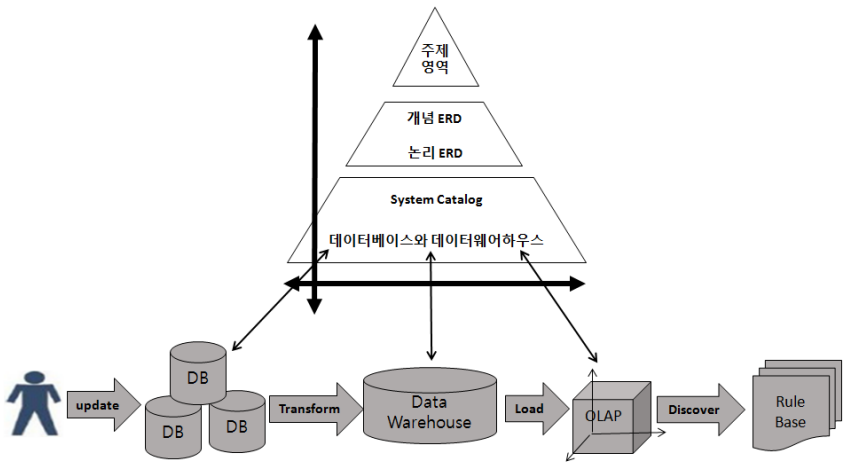
Ⅳ. 데이터 자원 관리의 변화

4.1 전통적인 환경하의 데이터 자원 관리

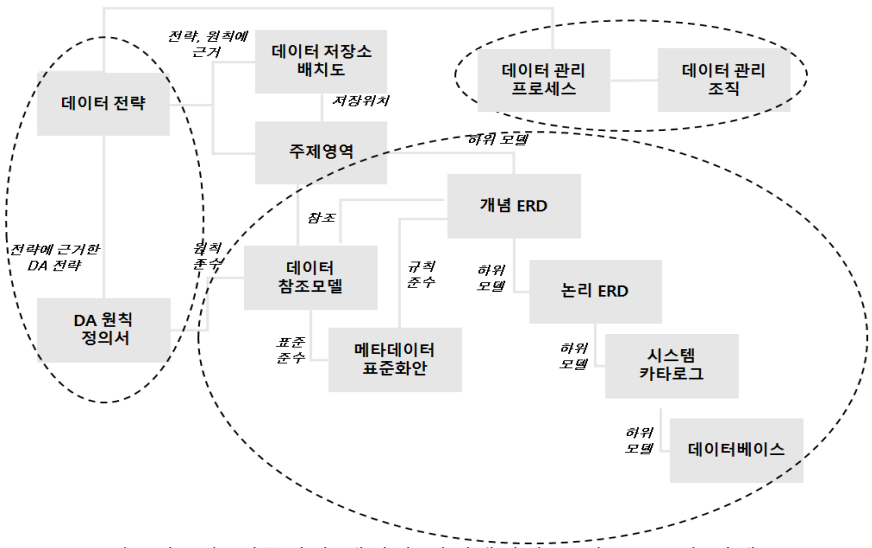
전통적 환경하의 데이터 생명주기를 고려할 때, 실세계에서는 데이터와 정보 개념은 활성화되었으나 지식 개념은 활성화되지 못 하였다. 대부분 기업들은 관계형 DBMS, 데이터 웨어하우스, ETL, SQL OLAP 등의 기술을 활용하여 시스템을 구성하였다. 사용자가 업무를 처리하면서 DBMS를 활용하여 데이터를 갱신한다. 갱신된 데이터는 주기적으로 데이터 웨어하우스에 축적된다. 사용자는 축적된 데이터를 근거로 다차원 분석을 하거나 통계 분석하여 지식을 도출하였다. 따라서 전통적인 시스템 구조에서는 데이터베이스와 데이터 웨어하우스가 핵심 기술이다.

시스템 아키텍처를 지원하기 위한 데이터 아키텍처는 이미 일반화되었다. 제Ⅱ장에서 언급한 것처럼, 데이터 아키텍처는 최상위의 개괄적인 수준에서부터 데이터베이스 수준까지 데이터에 관한 모든 구조를 통합하여 연계하고, 업무 및 기술들 타 아키텍처와의 전체적인 관계를 정립한다. 위 계층의 데이터 모델인 개괄 데이터 모델은 주제영역(Subject Area)으로 관리된다. 중간 계층의 데이터 모델인 개념데이터 모델과 논리데이터 모델은 개체관계도(Entity-Relationship Diagram)로 관리되어진다. 최하위 계층의 데이터 모델은 DBMS를 활용된다. 즉, 시스템카타로그로 물리 데이터베이스를 정의하고 실제 데이터는 SQL 언어로 관리된다.

전통적인 데이터 아키텍처의 구성요소를 보면 크게 3가지 영역으로 분류할 수 있다.즉, 전략 영역, 모델링 영역, 거버넌스 영역으로 분류할 수 있다. 모델링 영역은 주제영역, 개념 ERD, 논리 ERD, 시스템카타로그, 데이터베이스의 연계를 보여준다.



〈그림 12〉 전통적 환경하의 시스템 구조와 데이터 아키텍처



〈그림 13〉 전통적인 데이터 아키텍처의 구성요소들의 관계

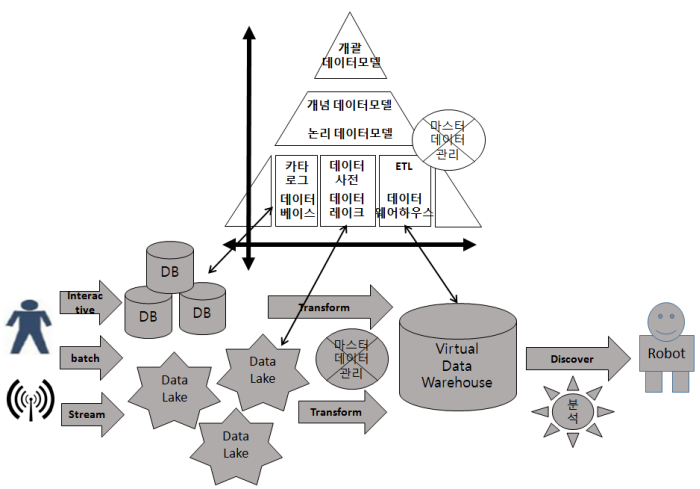
4.2 빅데이터 환경하의 데이터 자원 관리

빅데이터 환경하의 데이터 생명주기를 고려할 때, 지속적인 분석을 할 수 있는 구조적인 데이터와 비구조적인 데이터의 연계가 가장 중요하다. 대부분 기업들은 보유데이터 특성, 분석활용 목적 등을 고려하여 No SQL, HADOOP, Streaming 등 빅데이터 기술을 접목하여 다양한 형태로 데이터자원을 관리하고 있다.

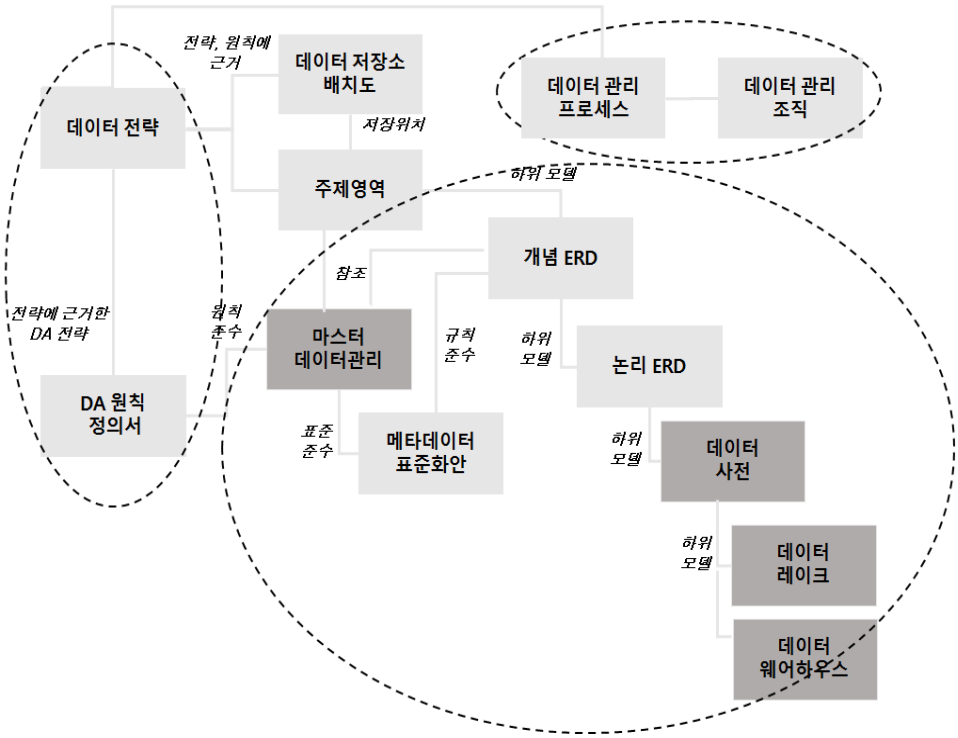
구조적인 데이터는 기존 시스템 구조를 활용할 수 있다. 반면에 비구조적인 데이터는 다양한 구조를 갖고 있기 때문에 하나의 구조적인 데이터 웨어하우스로 저장할 수 없다. 비구조적인 데이터는 데이터 레이크에 원래의 데이터 형태로 축적되어진다. 데이터 레이크에서는 비구조적인 데이터를 관리하기 위하여 데이터사전을 구성한다. 마스터 데이터 관리는 구조적인 데이터이든 비구조적인 데이터이든 모든 데이터 영역의 핵심데이터를 관리한다. 가상 데이터 웨어하우스는 전처리 과정을 통해서 구조데이터로 관리하면서 데이터 레이크의 비구조데이터도 가상적으로 연계할 수 있다. 따라서 빅데이터 환경하의 시스템 구조에서는 데이터 레이크와 마스터 데이터 관리가 핵심기술이다.

데이터레이크를 데이터웨어하우스와 비교해 보면, 데이터웨어하우스는 현업실무자가 직접 활용하기 위한 가공된 데이터 창고라면 데이터 레이크는 데이터분석자를 위해 가공 전의 원래 데이터 형태를 그대로 유지한다. 데이터웨어하우스는 스키마에 따라 데이터 업로드가 된다면, 데이터레이크는 스키마에 따라 데이터를 검색한다. 따라서 데이터레이크는 하둡 구조로 저장되고 관리된다.

빅데이터 환경을 지원하기 위한 데이터 아키텍처는 아직 정립되어 있지 않다. <그림 14>에 제시된 시스템 구조를 기준으로 데이터 아키텍처를 도출할 수 있다. 전통적인 데이터 아키텍처와 비교하여 보면, 개괄데이터 모델과 개념데이터 모델은 동일하다. 논리 데이터 모델은 비구조화된 원천데이터를 표시한다는 점에서 다르다. 최하위 계층의 데이터 모델은 데이터베이스, 데이터 레이크, 데이터 웨어하우스를 모두 관리해야 하기 때문에 기존 데이터 모델과 완전히 상이하다. 데이터베이스는 시스템카타로그로 관리하고, 데이터 레이크는 데이터사전으로 관리하고 데이터 웨어하우스는 ETL로 관리하면서 모든 데이터를 연계하기 위하여 마스터 데이터 관리가 존재한다. 구조적인 데이터이든, 비구조적인 데이터이든 마



〈그림 14〉 빅데이터 환경하의 시스템 구조와 데이터 아키텍처



〈그림 15〉 빅데이터 환경하의 데이터아키텍처 구성요소들의 관계

스터데이터를 통해서 연계되고 마스터데이터를 통해서 품질을 관리한다.

전통적인 데이터 아키텍처 구성요소와 비교해보면, 빅데이터 데이터 아키텍처의 구성요소는 마스터 데이터 관리, 데이터 레이크, 데이터 사전 등이 추가된다. 데이터 레이크는 구조화되지 않은 원천데이터를 관리한다. 데이터 사전은 비구조화된 데이터를 구조화된 데이터로 전개하는 중간자 역할을 수행한다. 마스터 데이터 관리는 데이터의 구조에 상관없이 전사 데이터를 연계해주는 역할을 한다.

Ⅴ. 결 론

본 논문에서는 전통적인 데이터 자원 관리와 비교하여 빅데이터 환경을 위한 데이터 자원 관리를 연구하였다. 본 연구의 의의는 크게 두 가지로 나누어 볼 수 있다.

첫 번째로 기존 연구는 빅데이터 환경을 위한 시스템 구조를 제안하였다면, 본 연구는 빅데이터 환경을 위한 데이터 아키텍처를 제안했다는 점이다. 과거에 기업의 일부 비즈니스 영역에서 빅데이터 분석이 수행되었다면, 앞으로는 기업 전사 비즈니스 영역에서 빅데이터 분석이 이루어지기 때문에 전사 빅데이터를 위한 데이터 아키텍처는 매우 중요해질 것이다.

두 번째로 본 연구에서는 빅데이터 아키텍처를 위한 주요 구성요소를 제안하고 그들 간의 관계를 제안하고자 한다. 기존 연구는 각 구성요소의 개념을 설명하였으나, 본 연구는 각 구성요소의 관계를 보여주고 있다. 특히 데이터 레이크나 마스터 데이터 관리가 타 구성요소와 연계 관계를 보여주었다는 점에서 기업 실무자에게 실질적인 도움이 될 수 있다.

참 고 문 헌

[1] 김승현, 박주석, 박재홍, 김인현, “빅데이터 환경에서 분석자원이 기업성과에 미치는 영향”, 한국빅데이터학회, 제1권, 제1호, 2016.

[2] 박주석, “데이터 중심의 공공 정보자원관리”, 한국정보화진흥원, 연구보고서, 2016.

[3] 장동인, “빅데이터로 일하는 기술”, 한빛미디어, 2014.

[4] 김인현 외 다수, “2015 한국데이터산업 백서”, 한국데이터진흥원, 2015.

[5] Carl Anderson, “Creating a Data-Driven Organization”, *O’REILLY*, 2015.

[6] Chen, H., R.H.L. Chiang, and V.C. Storey, “Business Intelligence and Analytics: From Big Data to Big Impact”, *MIS Quarterly*, Vol.36, No.4, pp.1165-1188, 2012.

[7] Mckinsey and Company, “Big Data: The Next Frontier for Innovation, Competition and Productivity”, *McKinsey Global Institute*, 2011.

저 자 소 개

박 주 석(Jooseok Park)

- 1981년 : 서울대학교 산업공학 (학사)
- 1983년 : 한국과학기술원 산업공학 (석사)
- 1990년 : University of California, Berkeley MIS(박사)
- 현재 : 경희대학교 경영대학 교수
- 관심분야 : 데이터베이스, 모델링, 아키텍처, 정보화전략 등

김 인 현(Inhyun Kim)

- 1982년 : 한국외국어 대학교 무역학과 (학사)
- 1984년 : 서울대학교 대학원 경영학과 (석사)
- 2011년 : 경희대학교 박사과정 수료
- 현재 : 투이컨설팅 대표이사
- 관심분야 : 데이터분석, 디지털혁신, 핀테크, 정보전략계획 등