# 빅 데이터와 산업체 활용
# Big Data and Industrial Applications

2015. 10.

한국과학기술정보연구원

## 머리말

지금 우리는 정보화 혁명에 이어 '데이터 혁명'이라는 새로운 물결에 직면하고 있다. 이는 정보화 혁명에서 창의적으로 발현된 스마트 기술로 인간 중심의 창조적 가치를 실현하는 시대가 오고 있음을 의미한다. 이러한 혁명의 시대를 이끌어갈 핵심 키워드로 '빅 데이터'가 새롭게 등장하여 빠른 속도로 확산되고 있다.

현재 세계 주요 연구기관에서는 빅 데이터의 사회경제적 파급효과에 대한 활발한 연구 및 논의가 진행되고 있다. 특히, 산업계에서의 빅 데이터 활용을 통한 그 파급효과는 참으로 높다할 수 있다. 아울러 과학기술 분야에서도 데이터 중심(Data Intensive)의 과학 패러다임 변화가 역동적으로 진행되고 있다.

맥킨지는 의료, 공공행정, 소매, 제조, 개인정보 등 다양한 부문에 빅 데이터를 적용할 수 있으며, 미국의 경우 최대 7천 억 달러의 경제적 효과가 창출될 것으로 예측하였다. 영국의 경제비즈니스 연구센터는 향후 5년 간 빅 데이터 관련 비즈니즈가 자국 내에 약 360조 원 이상의 경제적 파급효과를 미칠 것으로 전망하였다. 가트너는 2015년까지 전 세계적으로 440만 개, 미국 내에서는 190만 개의 IT 일자리가 창출될 것이라 전망하였다.

사회경제뿐만 아니라 과학기술 분야에서 혁신적 역량을 발휘하는 빅 데이터는 우리가 살고 있는 제반 사회문제의 근본적인 해결책을 제시하고 있다. 예를 들어 사회범죄, 테러, 재난재해, 저출산 고령화 등 지금 우리가 당면하고 있는 사회 현안들에 대하여 데이터 분석으로 최적의 대응을 위한 통찰력을 제공할 수 있다.

이와 같이 빅 데이터는 더 나은 사회를 과학적으로 디자인해 저비용 고효율의 삶을 가능케 하는 최고의 방법이 될 것이다.

이에 이 보고서가 향후 빅 데이터를 바탕으로 산업 기술 혁신을 가속화하기 위한 기초 자료로 활용될 수 있도록 산업체에서의 빅 데이터의 진정한 의미와 핵심 기술 그리고 현재 빅 데이터가 적용되고 있는 사례들을 정리하였다.

2015년 10월

한국과학기술정보연구원 이 상 민

목차





# 산업체에서의 빅 데이터의 중요성

## 빅 데이터의 정의

빅 데이터(Big Data)를 의미대로 해석하면 거대한 데이터이다. 기업의 관리하고 있는 대규모 DWH(Data WareHouse)의 용량은 1테라바이트 규모 정도이지만, 1,000 테라바이트(1페타바이트) 규모의 시스템을 빅 데이터라고 할 수 있을까?

최근 빅 데이터를 Facebook 등과 같은 쇼셜 네트워크에서 발생되는 데이터를 얘기하는 경우가 많다. 물론 Facebook의 회원수가 8억 명을 넘고, 여기에서 발생되는 데이터가 하루 10테라바이트 이상이 된다. 그리고 이러한 쇼셜 미디어는 RDBMS(관계형 데이터베이스)가 아니고, NoSQL(Not Only Structured Query Language)이라는 RDBMS(Relational Data Base Management System)와는 다른 데이터 관리 소프트웨어를 사용하고 있어, 빅 데이터를 NoSQL시스템이라고 하는 것인가?

이와 같은 현상을 감안하면 빅 데이터를 "1 페타바이트와 같은 매우 많은 양의 데이터를 NoSQL을 이용하여 처리하는 시스템"이라고 정의할 수 있지만, 과연 정확한 것인가? 그러나 이 정의 또한 정확하다고 할 수 없으며, 현재 이슈가 되고 있는 빅 데이터가 나타내는 의미는 좀 더 복잡하고 다양한 의미를 가지고 있다.

그 이유는 이미 RDBMS를 이용하여 대량의 데이터를 이용하고 있는 기업은 세계에 이미 존재하고 있기 때문이다. 예를 들어, Bank of America는 1.5 페타바이트 이상의 DWH을 보유하고 있으며, 세계적인 슈퍼마켓 체인인 월마트 스토어는 2.5 페타바이트 이상, 인터넷 경매 사이트 인 eBay는 6 페타바이트 이상의 데이터를 저장한 DWH를 실행하고 있다. 따라서 데이터의 양이 페타바이트 급이 된다고 하여 빅 데이터라고 할 수 없다. 테라바이트급 이상의 거대한 DWH시스템은 EDW(Enterprise Data Warehouse)라고 하며, 그 데이터베이스는 VLDB(Very Large Database)라고 한다.

NoSQL은 RDBMS에 비해 스케일 업(처리 성능을 향상시키기 위해 서버와 스토리지의 대수를 늘려 처리 성능을 향상 시키는 것)에 적합하다고 알려져 있다. 하지만 그렇다고 해서 RDBMS가 필요 없게 되는 것이 아니다. NoSQL은 문서나 이미지 등의 비정형 데이터의 처리에는 적합하지만, 숫자 등의 구조화된 데이터, 특히 데이터의 정확성을 중요시하는 처리 에는 적합하지 않다. 실제로 Facebook에서도 모든 작업을 NoSQL로 실시하고 있는 것이 아니라 RDBMS도 이용하고 있다. 즉, 데이터의 종류와 필요한 작업에 따라 RDBMS와 NoSQL을 병행하여 사용하고 있다.

현재 빅 데이터에 대한 명확한 정의는 없지만, 빅 데이터를 인터넷의 보급과 IT 기술의 진화에 의해 발생되는 지금까지 기업이 관리하여 온 데이터보다 더 고용량의 다양한 데이터를 관리・활용하기 위한 새로운 정보구조를 나타내는 것이라 할 수 있으며, 그 특성은 데이터 양, 속도(업데이트 속도), 다양성(데이터의 종류)로 표현할 수 있다.

빅 데이터는 현재 기업에서 활용하는 데이터 관리법인 DWH 및 OLTP(On-Line Transaction Processing)와 같은 시스템에 비해 다음의 세 부분에서 차이가 있다. 첫 번째는 데이터의 양이 많다는 것, 두 번째는 데이터의 종류가 많다는 것, 그리고 세 번째는 데이터의 변화하는 빈도가 높다(업데이트의 속도가 빠르다)는 것이다. 따라서 현재 발생되는 데이터가 이러한 조건을 가지고 있어 종래의 데이터관리시스템에서는 효율적인 처리가 어려운 데이터를 처리하는 시스템을 빅 데이터라고 한다.

빅 데이터의 특징은 데이터의 양뿐만 아니라 데이터의 종류가 매우 다양하다는 것이다. 데이터의 종류에는 회계정보시스템, ERP시스템 등과 같은 기업의 각종 기간 시스템에서 생성되는 숫자나 문자열로 구성되는 구조화된 데이터뿐만 아니라 문장, 음성, 동영상 등 멀티미디어 데이터 등의 비정형 데이터가 포함된다. 또한 e-메일 데이터와 XML 데이터와 같은 반구조화 된 데이터, 또한 각종 센서 및 장치에서 나오는 데이터 및 통신 로그 등과 같은 발생 빈도가 매우 많은 데이터도 포함된다. 또한 이러한 데이터는 사내뿐만 아니라 인터넷상의 외부에 있는 경우가 많다.

따라서 대상으로 하는 데이터 종류의 차이가 기존의 시스템과 빅 데이터를 구분하는 하나의 팁이 될 수 있다. 현재 빅 데이터의 활용을 선도하고 있는 기업의 대부분은 Google과 Facebook과 같은 Web 서비스 사업자이다. 그리고 이들 기업의 데이터 활용은 기존의 판매 데이터와 고객 데이터와 같은 기업 내부에 존재하는 데이터가 아닌 Web 상에 존재하는 문장이나 이미지 같은 데이터가 중심이 되어 있다.

그리고 현재까지 많은 기업들이 데이터의 활용에 있어 고객 데이터를 대상으로 한다고 하여도, 활용 목적이 개별 고객의 특성이 아니라 전체적 관점에서 데이터를 집계하여 획득되는 경향(동향) 중심 정보가 대상이었다. 그러나 빅 데이터 관점에서 데이터를 처리하는 Web 서비스 사업자는 고객의 개별 특성을 파악해 Amazon에서 제공하는 제품 추천과 같은 세분화된 정보를 이용한다. 그리고 크게 다른 점은 데이터 처리에 대한 정확도보다 속도가 더욱 중시되고 있는 점이다.

기존의 데이터는 RDBMS가 중심이라면, 빅 데이터는 NoSQL가 기본이 된다. 기존의 DWH는 대용량인 경우에도 구조적 데이터 중심으로 데이터 갱신 빈도도 월 단위로 변화 빈도도 그다지 높지 않았다. 따라서 "대용량 +비정형 데이터 +업데이트 속도"에 대응하는 새로운 정보체계로 빅 데이터가 주목받게 된 것이다.

## 빅 데이터가 주목받는 이유

최근 기업 비즈니스에 빅 데이터의 활용이 이슈가 되고 있다. 여기서는 기업 비즈니스에 빅 데이터를 활용이 필요한 이유는 다음과 같다.

### *정확한 현황 파악과 예측의 중요성(보다 폭 넓은 정보의 활용)*

현재 사회 및 경영환경은 그 변화의 폭이 점점 심해지고 미래의 예측이 더욱 더 어려워지고 있다. 이러한 환경에서 기업이 충분한 경쟁력과 수익을 유지하기 위해서는 외부의 변화에 민첩하게 대응하는 것이 필수가 된다. 이렇게 민첩하고 신속한 대응을 위해서는 정확하고 신속한 의사 결정이 필수적이지만, 이를 위해서는 먼저 현재의 상황, 그리고 어떤 일이 일어나고 있는지를 파악하고, 미래에 무엇이 일어나는지를 정확하게 예측해야 한다.

따라서 기존 기업이 보유한 데이터에 근거하여 의사결정을 수행하였지만, 기업 내부 데이터뿐만 아니라 기업 외부 데이터까지 활용의 폭을 확대하여, 이를 통해 외부 환경까지 고려한 정확한 파악과 예측이 요구되고 있기 때문인 것이다.

이러한 정보를 바탕으로 정확하고 적절한 의사결정을 수행하여, 급변하는 경영환경에 대응하는 것이 기업과 사회에 점점 강하게 요구되고 있다.


그림 1. 빅 데이터의 적용 배경 – 정확한 상황 파악 및 예측의 중요성

### *취급 정보량의 폭발적인 증가*

두 번째 이유는 기업 활동 및 사회 활동에 검토 대상이 되는 정보 자체가 폭발적으로 증가하고 있다는 것이다.



사업 및 기업 활동에 곳곳에서 크고 작은 컴퓨터와 센서(디바이스)가 설치되고, 네트워크를 통하여 활동에 관련된 다양하고 대량의 데이터가 수집, 축적되는 "정보 폭발"이라고 할 수 있는 현상이 발생하고 있다. 그리고 기업에서도 대상으로 할 정보가 기업 사내 활동으로 생성된 정보뿐만 아니라, 기업 외부에서도 대량으로 유입되고 있는 시대가 되고 있다.

전 세계에서 1년간 발생하는 정보(데이터)의 양이 현재는 엑사바이트(Exabyte)를 초과하여 제타바이트(Zetabyte)의 영역에 이르고 있으며, 현재 존재하는 디지털 데이터의 약 80 %가 지난 2년 동안 발생하였다. 기업 활동에 대상으로 해야 할 정보의 폭이 확대된 것뿐만 아니라, 정보의 양 또한 압도적으로 증가하고 있는 것이다.


그림 2. 빅 데이터의 적용 배경 – 정보량의 폭발적 증가

### *정보 처리 비용의 저가격화*

대상으로 할 데이터가 대량으로 존재한다고 해도 데이터에 그 자체만으로는 대단한 가치가 없다. 존재하는 대량의 데이터 중에서 기업 비즈니스에 필요한 정보를 얻고, 정확한 의사결정을 통하여 기업 활동에 적용하는 것을 통하여 경제적인 가치가 발생하는 것이다.

기업 활동에 적용하기 위한 의사결정에 소요되는 시간을 어떻게 단축하는가가 중요하며, 이를 위해서는 빠르고 거대한 컴퓨터 작업이 필요하다. 또한 그 처리 비용이 적정 범위로 유지되지 않으면, 정보 활용의 장점이 상쇄되어 버릴 가능성도 있다.

다행히 최근의 컴퓨터 처리 성능은 10년과 비교하여 약 220배 증가하였으며, 단위 성능 당 비용은 100분 의 1 정도까지 낮아져, 대량의 데이터를 적정한 비용으로 처리하기 위한 인프라는 정비되었다고 할 수 있다.

따라서 정확한 의사 결정을 신속하게 수행하기 위해서는 보다 다양한 정보를 활용해야 하며, 그 정보의 양이 급증하고 있지만 이들 정보를 신속하게 처리하기 위한 인프라도 적정한 비용으로 활용 가능한 시대가 되었다.

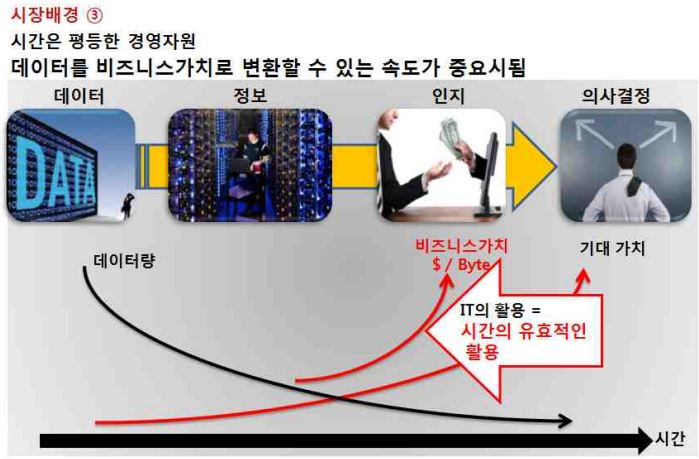



그림 3. 빅 데이터의 적용 배경 – 정보처리 비용의 저가격화



## 빅 데이터가 의미하는 것

빅 데이터를 "기술"과 "비즈니스"의 두 가지 관점에서 파악하면 다음과 같다.

***다른 구조의 데이터를 포괄적으로 취급***

기업 활동에 활용되는 데이터는 2차원 테이블 구조에 저장되는 구조화된 데이터와 각종 센서 디바이스 등에서 수집되는 준구조화된 데이터(반구조화된 데이터), 그리고 2차원 테이블 구조에 저장이 어려운 비정형 데이터 등이 있다.

이와 같이 데이터 구조의 관점에서 보면 데이터의 종류는 다양하지만 비즈니스 가치의 원천으로 보면 기업에게 동일한 가치를 가진다. 기업에서는 데이터를 통합하여 포괄적으로 관리하고 싶지만 데이터 구조가 다르면 관리시스템을 구축하는 것은 어려운 일이었다. 이러한 다양한 데이터를 포괄적으로 운영 관리하는 것을 빅 데이터라고 한다.

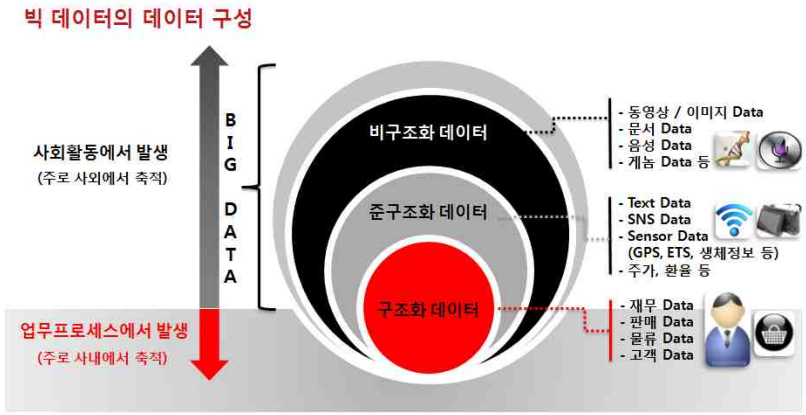



그림 4. 빅 데이터의 데이터 구성

***내외부에서 발생하는 데이터를 포괄적으로 취급***

지금까지 많은 기업들은 경영활동에 활용하는 정보는 주로 자사의 비즈니스 프로세스(생산 공정 및 영업 관리 프로세스)에서 발생한 정보가 대상이었다. 그러나 현재는 각종 컴퓨터 장치와 인터넷이 보급되어, 이들을 이용한 다양한 활동에서 데이터가 폭발적으로 증가하고 있어, 이러한 데이터를 내부 비즈니스 프로세스에서 발생한 데이터와 포괄적으로 취급하여 새로운 비즈니스 기회, 새로운 가치를 모색하는 활동이 활발해지고 있으며, 이러한 활동이 빅 데이터 활동이라고 한다.

## 빅 데이터의 비즈니스 적용 방법: 빅 데이터 관리와 빠른 데이터 처리 (Fast Data Processing)

빅 데이터 그 자체로만으로는 가치가 있는 경우는 드물며 일반적으로 이를 수집/관리, 활용하여 의사 결정에 적용하는 "빅 데이터 관리(Big Data Management)"라는 활동을 통하여 그 가치가 발휘된다.

즉 빅 데이터 관리에 있어 정보를 수집하고 집계/분석하여, 이를 기업 활동에 어떻게 활용하고 어떤 이익을 얻을 수 있는가 하는 비즈니스 모델 수립이 매우 중요한 것이다.

빅 데이터 관리에는 다음과 같이 크게 2 개의 응용 패턴으로 나눌 수 있다:

먼저 다양한 데이터를 수집하고 여기에 분석을 통하여 대상으로 하는 사건에 일정한 규칙성을 발견하여 비즈니스에 활용 하는 것이다. 예를 들면 빅 데이트 성공 사례로 거론되는 "슈퍼에서 유아 기저귀 옆에 맥주를 진열 해 두면 기저귀를 사러 온 아버지가 함께 구입해 주는 경우가 많다"라는 규칙성을 발견한 사례가 있다. 이러한 규칙성을 발견하는 것은 기존의 데이터 분석에서는 매우 어려운 일이지만, 데이터 분석 기술의 발전과 컴퓨터 처리 능력 향상으로 현재는 실현 가능하게 되고 있다.


그림 5. 빅 데이터 관리

다른 하나는 비즈니스 기회를 놓치지 않기 위한 것이다. 네트워크에서 유통되는 다양한 데이터를 포착하고 미리 예측한 비즈니스 패턴이나 규칙성에 일치하는 데이터의 흐름이 인식되었을 경우에는 미리 규정된 처리를 수행한다. 이것을 최근에는



"패스트 데이터 프로세싱(Fast Data Processing)"이라 하며, 그 활용 예로는 주식 알고리즘 트레이딩이 있다. 이것은 주가의 동향을 지속적으로 감시하여, 규정된 패턴과 일치하면 이에 해당하는 적절한 조치를 자동으로 수행하여 수익성을 높이거나 리스크를 감소시키는 것이다. 패스트 데이터 처리는 규칙성을 발견하는 작업과 그 규칙성에 따라 적절한 실행을 수행하는 것이 관건이다.

## 기업에서의 빅 데이터의 세 가지 가치

사회에서 발생되는 모든 일을 데이터로 파악될 수 있는 현재의 경영환경에서 새로운 시장 및 비즈니스 기회, 가시화되지 않는 고객의 요구사항, 미래를 위한 추진 전략 수립에 데이터의 이용 및 활용은 필수적으로 요구되고 있다. 즉, 빅 데이터로부터 파생되는 가치에 따라 향후 기업의 경쟁 우위가 좌우된다고 할 수 있다.

***보이지 않는 것을 시각화한다.***

제품, 서비스, 고객의 각각의 단위에서 상세한 데이터를 수집/처리하여 자사가 보유한 데이터와 이업종 데이터와 조합하여 분석함으로서 기존에 파악되지 않았던 복잡한 세계를 시각화・계량화할 수 있다.

***새로운 관점에서 가치 창출이 가능하게 된다.***

시각화된 결과에 대하여 원인 분석 및 수학적 모형화를 통해 새로운 비즈니스 가치를 창출하고 새로운 비즈니스 기회 확대 및 기업의 업무 생산성 향상을 도모할 수 있다.

***현재의 상황에서 미래를 예측할 수 있다.***

현황 분석에 실시간 처리 기술, 수요 예측 등 미래 예측 기법을 활용하여 새로운 비즈니스 기회를 확대할 수 있다.

## 빅 데이터의 활용 범위

현재 기업에서 다양한 소프트웨어 및 하드웨어 등의 IT 기술이 사용되고 있지만, 이러한 IT 기술은 기업별로 차별화 수 없다고 한다. 이것은 기업이 활용하는 IT 기술은 대금을 지불하면 구축이 가능하기 때문에, 그것만으로는 차별화할 수 없다는 것을 의미한다. 그러나 데이터와 정보는 단순히 자금을 투자해도 해결할 수 있는 문제는 아니다. 데이터 및 정보 등의 대부분은 실제 활동에 의해 만들어진 결과이며, 기업의 노하우이다. 즉 데이터와 정보는 3M(사람, 물건, 돈)과 동등한 가치를 가진 경영 자산으로 인식되고 있다.

따라서 새로운 정보자산의 활용이라는 측면에서 보면 기업에서 빅 데이터의 활용은 당연 한 결과이지만, 빠른 하드웨어 및 고급 소프트웨어를 사용하는 것이 아니라 데이터 및 정보를 비즈니스에 활용하는 것, 즉 데이터 및 정보라는 자산 활용이라는 측면에서 검토하면 데이터의 활용 범위는 무궁무진하다고 할 수 있다.

현재 기업 비즈니스에 빅 데이터를 활용하고 있는 사례는 다음과 같다.

첫째로, 현재 빅 데이터를 가장 많이 활용하고 있는 기업은 Web 서비스 사업자이다. 예를 들면 Google은 검색 및 무료 응용 프로그램에서 축적한 방대한 데이터를 기반으로 광고 사업에 적용하고 있다. 또한 Facebook과 같은 소셜 미디어도 방대한 회원 데이터를 기반으로 광고와 게임 등의 소프트웨어 판매 등으로 수익을 올리고 있다. 또한 Amazon과 같은 인터넷 쇼핑몰에서는 회원 데이터, 구매 내역, 클릭 스트림(사이트 내에서 고객의 움직임) 등의 데이터를 사용하여 과거 이력이나 제품 추천을 제시함으로써 회원 고객별로 차별화된 개별적으로 서비스를 제공하여 구매 의욕을 높이는 정보 제공을 실시하고 있다.

빅 데이터의 이용은 Web 사업자와 EC뿐만 아니라, 통신사업자가 휴대 전화 등의 통신 로그를 분석함으로써 해당 고객의 통화 패턴 및 통신 상대자의 통신사업자가 어디가 많은 지를 분석하여, 통신사업자 변경 가능성을 사전에 감지하여 개별적 마케팅을 실시하여 이탈을 방지하거나, 반대로 친구 추천 캠페인 등 마케팅 등에는 이미 활용하고 있다. 이렇게 Web 사업 및 통신 사업자는 디지털화된 데이터와 정보를 관리하고 있어 간단하게 정보 활용이 가능하겠지만, 그렇지 않은 기업에서는 빅 데이터 활용을 어떻게 가능할 것인가?

다음 사례를 통하여 그 적용 방법에 대하여 검토해 보자.

손해보험회사가 자동차 내비게이션의 GPS에서 계약자의 운전 상황에 대한 데이터를 수집하면 연령, 마일리지, 면허의 종류 등의 정보뿐만 아니라 가입자당 실제 주행과 운전 상황이 파악이 가능하게 되어 계약자 개인별로 위험을 분석하여 기업의 마진의 확보와 가입자 의 가격 만족도를 양립 가능하게 할 수 있다. 또는 신용카드회사가 카드가 이용되는 장소 와 이용자의 스마트 폰의 GPS 데이터를 조합하여 카드의 부정사용을 사전에 탐지하는 일도 가능하게 할 수 있다.

금융회사, 통신사업자 등과 같은 서비스 제공 기업 외에 실제 상품을 개발하여 판매하는 기업에서는 빅 데이터 활용을 어떻게 해야 할 것인지가 문제가 될 수 있다. 그러나 어느 기업에서도 판매 행위와 고객은 존재하고 있기 때문에, 제품에 대한 고객의 의견과 고객의 요구를 파악하는 것은 중요한 문제이다. 이미 소매업을 중심으로 실시되고 있는 소셜 미디어 의 코멘트 정보에서 자사나 자사의 상품에 관한 의견을 파악하여 마케팅 시책에 이용하며, 상품 기획 및 개발에 활용하는 것은 모든 업종에서 빅 데이터를 활용하는 공통된 하나의 방법이 될 수 있다.

또한 빅 데이터는 사회 인프라 및 1 차 산업 등에서의 이용이 가능하다. 예를 들면, 도로 에 설치되어있는 센서, 자동차에 설치된 교통카드 이용정보 및 GPS 데이터를 이용하여 수집되는 교통량 데이터와 신호등 제어를 연동하면 혼잡 완화 및 이동 시간의 단축, 그리고 CO2 배출량의 저감이 실현 가능하게 된다. 각 병원에 보관되어있는 의료 기록 및 투약 정보와 다양한 검사 데이터를 통합 관리함으로써 의료비용의 절감, 의료 사고의 사전 예방, 원격 진료의 보급 촉진 등을 도모할 수도 있다.

또한 IT화가 지체된 1차 산업의 효율화에도 유효하게 적용 가능하다. 예를 들면 논밭 에 기상 센서를 설치하여 기상 데이터와 수확량, 품질 등의 데이터간의 관계를 파악하여 농업 프로세스를 최적화하여 생산성과 수익성 향상에 기여 가능할 것이다.

빅 데이터를 어떻게 활용하고 가치의 창출 여부는 기업의 아이디어와 노하우 또는 사업 전략 그 자체가 되며, 기업이 향후 추진해야 할 과제라고 할 수 있다. 그림 6에 업종별로 이미 빅 데이터를 활용된 사례를 정리하였다.


그림 6. 업종 별 빅 데이터의 활용 현황

그림 6에서 제시한 활용 사례에는 이미 기존 시스템에서 수행 되어 온 것도 많지만 빅 데이터를 활용한 시스템과의 차이는 사용하는 데이터의 종류의 다양성에 있다.

앞서 언급한 신용카드의 부정사용을 감지하는 경우에 지금까지는 카드가 사용된 점포나 ATM 등의 장소와 시간, 계약자의 주소, 지금까지의 사용 내역 등을 참조하여 사용된 장소가 범위를 벗어나는 경우에 부정사용이 아닌가를 감지하지만, 예를

들면 계약자의 스마트 폰 GPS 데이터나 방범 카메라의 동영상 데이터 등을 같이 이용하면 기존의 방식보다 더 정확하고 신속하게 부정사용을 감지 할 수 있게 된다.

빅 데이터는 다양한 활용 범위를 가지고 있기 때문에, 비즈니스에 어떻게 활용하면 좋을까를 생각하는 것이 중요하다. 안이하게 소셜 미디어의 정보를 도입하여 마케팅에 적용하면 좋은 것이라고 생각하면 빅 데이터의 가치를 오인할 가능성이 높다. 중요한 것은 데이터를 자산으로 재인식하고 자산이라면 사장되어 있는 자산을 현재까지 간과하였던 자산을 찾아서 활용해야 한다는 구상이 중요하다.

그러나 빅 데이터를 활용하면 모든 것이 잘되는 것은 아니다. 비록 데이터를 분석하여 매출이 확대될 제품을 파악한다고 하여도 즉시에 시장에 공급하지 못하면, 아무 의미가 없는 것이다. 그리고 발생되는 이익보다 데이터 분석 비용이 과도하게 발생하면 본말이 전도되는 현상이 발생한다. 이러한 제품 라이프사이클상의 효율적 관리와 경제적인 데이터 분석 비용 구현이 빅 데이터 활용상의 하나의 과제라 할 수 있다.



## 기업에서 빅 데이터 활용을 위한 준비

최근 빅 데이터가 화제가 되고 있고, 기업 비즈니스에 적용을 해야 한다고 하고 있지만, 과연 빅 데이터가 실제로 잘 활용이 될 것인가에 대해서는 의문이 있을 수 있다. 그 이유로는 현재까지 데이터 분석을 기반으로 한 정보시스템, 예를 들면 BI(Business Intelligence)이나 지식 관리 등 다양한 시스템의 이용이 주장되었지만, 실질적으로 활용에 성공한 기업은 그리 많지 않기 때문이다.

다음 그림은 2010년 12월에 일본의 IT 전략 및 컨설팅 기업인 ITR Co.이 일본 내 기업을 대상으로 정보 분석 도구의 이용 현황을 조사한 결과이다.

이 조사 결과를 보면 불행히도 정보 분석 도구를 "효과적으로 이용하고 있으며 비즈니스에 적용하고 있다"고 회답한 기업은 전체의 8 % 밖에 되지 않고 있다.

그림 7는 ITR Co.이 2001년부터 매년 조사하고 있는 “국내 IT 투자 동향 조사” 결과 중, 주요 IT 동향의 하나인 “정보・지식의 공유/재사용 환경의 구축”의 실시율과 3년 후 실시 예정 비율을 나타낸 것이다.

그림 7에서 알 수 있듯이 정보・지식의 공유/재사용 환경은 불행히도 실시율은 30 % 정도이며 2005년부터 전혀 증가하지 않고 있다. 3년 후 실시 예정은 60% 이상으로 답변하고 있지만, 실질적인 운영은 30% 미만에 머물고 있는 상황이다. 기업에서 기간계 정보시스템의 도입은 촉진되고 있지만 정보・지식의 공유/재사용에 관련된 정보시스템의 도입은 도입 의지가 있어도 실질적으로 도입이 되지 않는 경우가 많다.

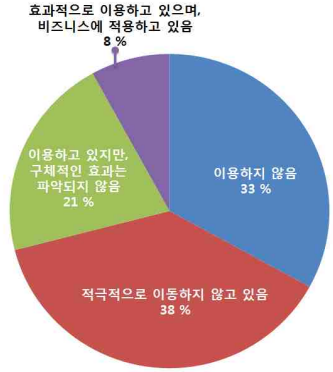



그림 7. 일본 기업의 정보 분석 도구의 사용 현황
(조사자: 일본 ITR Co., 조사 시기: 2010.12)

따라서 이와 같은 원인에 대하여 사전에 분석하는 것도 빅 데이터 도입 촉진을 위해서는 필요하다.

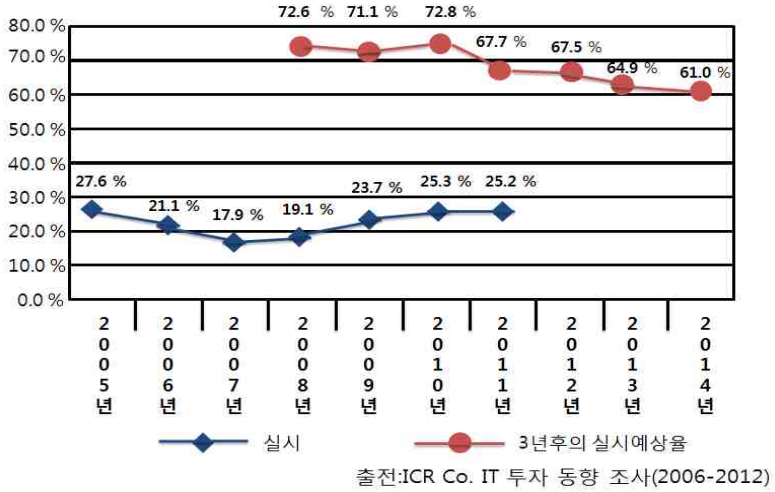



그림 8. 정보・지식의 공유/재사용 환경 실시 비율 추이

빅 데이터 활용의 본질은 데이터 분석을 비즈니스에 연결해 기업의 경쟁 우위를 획득하는 것이다. 이를 위해서는 기업은 어떤 환경을 구축해야 할 것인가?

기본적으로 다음의 세 가지 요소를 정비할 필요가 있다.

### *데이터 환경*

빅 데이터의 특징은 데이터의 종류가 다양성에 있으므로, 그 다양한 데이터에 대하여 다각도에서 분석 할 수 있는 데이터를 제공하는 것이 필요하다. 아무리 우수한 분석 도구와 알고리즘이 있어도 데이터가 없으면 실질적인 분석이 불가능하기 때문이다.

또한 데이터가 체계적으로 관리되고 있는 것도 중요하다. 어디에 어떤 데이터가 있고 그것에 대한 책임자는 누구이고, 어느 정도 정확한 데이터인지, 언제 생성된 데이터인지가 체계적으로 관리되지 있지 않으면 정리되지 않은 창고처럼 빅 데이터는 보물섬이 아니라 쓰레기 더미 되어 버린다.

### *도구 환경 정비*

쿼리 툴, 리포팅 툴, OLAP(On-Line Analytical Processing) 클라이언트 도구, 분석 도구 등 대상이 되는 데이터를 분석하기 위한 적절한 도구가 구비되어 있어야 한다. 그리고 아무리 고급 도구도 그것을 사용 가능한 인력이 없으면 무용지물이 되어 버리기 때문에, 분석 업무에 대한 교육 및 지원 센터 등의 배치도 중요하다 .



### *정보화 인지 능력*

여기서 말하는 정보화 인지 능력은 통계에 대한 지식과 도구의 사용법의 인지뿐만 아니라, 데이터 활용에 대한 아이디어와 데이터에 포함된 의미를 파악하는 능력이 포함된다.

그리고 사실 가장 중요하다고 것은 “데이터를 중시”하는 기업 풍토이다. 경영자, 비즈니스 부문의 사용자, IT 부서 등 전사적 차원에서 분석의 중요성을 이해 한 후, 각각의 입장 에서 세 가지 요소의 정비 및 능력 향상에 적극 노력하고 있는 상태를 구축해야 한다.

그러나 “데이터를 중시”하는 기업 문화의 구축이 가장 어려운 장벽이 될지도 모른다. 예를 들어, 데이터를 분석하고 업무에 반영하려고 해도 상사로부터 “데이터보다 현장을 중요시해라, 데이터 분석할 틈이 있으면 거래처나 현장에 나서라”라든지, “데이터 분석에 의한 방식은 지금까지 수행해 온 방식과 다르다”등과 같은 현재의 업무 방식의 전환하고자  하는 기업 풍토가 요구된다. 그리고 이러한 시책의 좋고 나쁨보다 그것이 누구의 의견인지, 발언권이 있는 사람의 의견만이 채택되는 기업 풍토는 지향될 필요가 있다.

물론, 상식과 경험에서 획득된 지식은 어떤 사실에 근거할 수 있지만, 그것은 과거의 것으로 현재는 다를 수가 있다. 따라서 데이터에서 확인해 보자는 기업 풍토가 없으면 빅 데이터를 충분히 활용하는 것은 어렵게 된다.

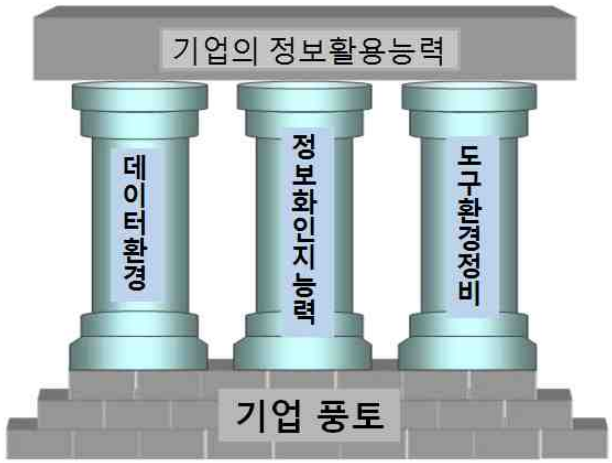



그림 9. 빅 데이터 환경 구축을 위한 세 가지 지원 요소

# 빅 데이터 컴퓨팅 기술

## 빅 데이터 등장 배경

1990년 이후 인터넷이 확산되면서 정형화된 형태의 데이터와 비정형화된 형태의 데이터가 무수히 발생하면서 정보 홍수(information overload) 개념이 등장했고, 이것이 오늘날 빅 데이터 개념으로 이어진 것이다. 개인화 서비스와 SNS(social network services)의 확산으로 기본 인터넷 서비스 환경이 재구성되었다. 검색과 포털[1] 위주였던 서비스를 제공하는 서비스가 통신, 게임, 음악, 검색, 쇼핑 등의 영역에서 개인화 서비스와 소셜 네트워크 서비스를 제공하는 환경으로 바뀌었다. 정보통신 기술(information&communication technology; ICT)시장조사 기관인 IDC(international data corporation) 디지털 유니버스(digital universe)가 조사한 보고서에 따르면 전 세계 디지털 데이터양[2]이 제타바이트 단위로 2년 마다 2배씩 증가해서 2020년에는 약 40제타바이트가 될 것이라 전망하고 있다. 40제타바이트는 전 세계 해변에 있는 모래알의 양인 7억 50만 조의 57배에 해당하는 숫자이며, 특히 스마트 폰의 보급으로 데이터가 매우 빠르게 축적되어 제타바이트 시대를 스마트 시대라고 한다.

데이터 양이 엄청나게 증가하여 기존의 데이터 저장, 관리, 분석 기법으로는 처리하는데 한계가 있어 정보 기술의 패러다임도 아래 표와 같이 바뀌었다. 그리고 이는 빅 데이터 용어를 등장시켰는데 패러다임이 지능화와 개인화된 시대를 빅 데이터 시대라고 한다.

표 1. 정보 기술의 패러다임 변화

| | PC 시대 | 인터넷 시대 | 모바일 시대 | 스마트 시대 |
|---|---|---|---|---|
| 패러다임 변화 | 디지털, 전산화 | 온라인화, 정보화 | 소셜화, 모바일화 | 지능화, 개인화, 사물정보화 |
| 정보 기술 이슈 | PC, PC통신, 데이터베이스 | 초고속 인터넷, www, 웹서버 | 모바일 인터넷, 스마트폰 | 빅데이터, 차세대 PC, 사물 네트워크 |
| 핵심 분야(서비스) | PC, OS | 포털, 검색 엔진, Web2.0 | 스마트폰, 웹 서비스, SNS | 미래 전망, 상황인식, 개인화 서비스 |
| 대표기업 | MS, IBM | 구글, 네이버, 유튜브 | 애플, 페이스북, 트위터 | 구글, 삼성, 애플, 페이스북, 트위터 |
| 정보 기술 비전 | 1인 PC | 클릭 e-Korea | 손 안의 PC, 소통 | IT everywhere, 신가치창출 |

1) 사이트(portal site)는 월드 와이드 웹에서 사용자들이 인터넷에 접속할 때 기본적으로 거쳐가도록 만들어진 사이트를 말함. 포털이라는 단어는 영어 낱말로서 정문, 입구를 뜻함. 사용자들이 필요로 하는 정보 또는 그에 대한 메타데이터를 종합적으로 제공함. 예를 들면 구글, 네이버, 야후 등

2) 1테라바이트(TeraByte; TB) = 1024GB, 1페타바이트(PetaByte; PB) = 1024TB, 1엑사바이트(ExaByte; EB) = 1024PB, 1제타바이트(ZetaByte; ZB) = 1024EB, 1요타바이트(YottaByte; YB) = 1024ZB

빅 데이터 개념이 등장하면서 데이터에 관심이 높아지고 있으며, 정보 통신 기술이 발전하면서 데이터도 규모, 유형, 특성에 따라 변화하고 있다. 아래 그림은 이런 데이터의 변화 방향을 나타낸 것이다. 특히 시스코(cisco)는 2012년 글로벌 모바일 데이터 트래픽[3] 전망 업데이트(global mobile data traffic forecast update)에서 2016년에는 세계 모바일 데이터 트래픽이 2011년 대비 18배 증가하여 10엑사바이트를 초과할 것이라고 전망하고 있다.

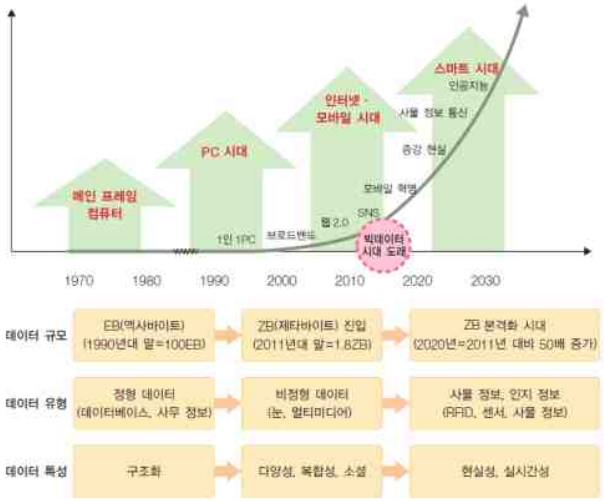



그림 10. 정보 통신 기술 발전에 따른 데이터의 변화 방향

빅 데이터를[4] 개인화 서비스 측면에서 생각해 보면, 고객의 성향이나 수입 규모, 소비 형태 등을 바탕으로 하는 개인화 서비스는 과거에도 있음. 신상품이 들어오면 고객의 취향에 맞춰 해당 상품 정보를 팸플릿(pamphlet)이나 휴대폰 문자 메시지로 고객에게 제공하는 것이 초기 형태의 빅 데이터 서비스이다. 이후 빅 데이터로 스마트 기기 사용자가 본 영화, 들은 음악, 찍은 사진, 촬영한 동영상, 쇼핑한 물건, 저녁을 먹은 레스토랑 등 모든 활동이 노출된다. 이런 수많은 비정형 데이터를 분석하여 개개인의 생각과 행동을 분석하고, 경향과 패턴을 파악할 수 있게 되었으며, 패턴 분석으로 대중의 변화를 예측하고 개인에게 최적화된 맞춤형 서비스까지 가능해졌다.

빅 데이터가 계속해서 차세대 이슈로 떠오르는 이유는 다음 세 가지로 요약할 수 있다:

3) 서버에서 전송되어지는 데이터의 총 전송량을 말함.
4) RFID(radio-frequency identification) 기술은 전파를 이용해 먼 거리에서 정보를 인식하는 기술을 말함. 예를 들면 교통카드임.



***정보 통신 기술의 주도권이 데이터로 이동***

모바일, 클라우드, 소셜 네트워크 서비스 등의 등장으로 정보 통신 기술의 주도권이 인프라와 기술 등에서 데이터로 이전되고 있으며, 이에 데이터의 폭발적인 증가에 대응하고 데이터를 분석하는 방법이 정보 통신 기술의 가장 중요한 이슈로 부각되어 빅 데이터를 정보 통신기술 시장과 기술 발전의 핵심 주제로 인식된다.

***공간, 시간, 관계, 세상 등을 담은 빅 데이터***

스마트 기기의 확산으로 사용자가 자발적으로 참여하고 정보를 생성하는 소셜 데이터 혁명이 발생했으며, 소셜 데이터 혁명은 정보의 생성자, 규모, 파급 효과 등에서 1990년대 기업이 고객의 정보를 축적했던 정보 혁명과는 구분한다.

페이스북, 트위터 등 소셜 네트워크 서비스 이용 확산과 소통 방식의 변화는 데이터 변혁을 가져오는 가장 중요한 요인이 되었다. 소셜 네트워크 서비스로 제공되는 정보는 지식 정보와 함께 정서적인 공감에 바탕을 둔 감성적 정보가 큰 비중을 차지하고, 소셜 네트워크 서비스에서는 개인의 취향이 더욱 직접적으로 반영되며, 진실성과 진정성, 관련성이 증가되어 데이터로서 가치가 매우 높다.

***빅 데이터는 미래 경쟁력과 가치 창출의 원천임***

빅 데이터에는 잠재적 가치와 위험이 공존하는데, 사회, 경제적으로 승패를 좌우하는 핵심 원천이 될 것으로 평가된다. 이에 세계 각국의 정부와 기업은 빅 데이터가 향후 기업의 승패를 가늠할 새로운 경제적 가치의 원천이 될 것이라 기대한다.

빅 데이터에서 유용한 정보를 찾고 잠재된 정보를 활용할 수 있는 기업이 경쟁에서 시장을 선도할 것으로 예상되어 맥킨지(mckinsey)[5], 이코노미스트(economist)[6], 가트너(gatner)[7] 등은 빅 데이터를 활용한 시장 변동 예측, 신산업 발굴 등 경제적 가치창출 사례 및 효과를 제시하고 있다.

5) 맥킨지는 시카고 대학 경영학부 교수이자 공인회계사인 제임스 맥킨지가 1926년에 설립한 컨설팅 회사임.
6) 이코노미스트는 영국에서 발행되는 국제 정치 경제 문화 주간지임.
7) 가트너는 미국의 정보 기술 연구 및 자문 회사임.

## 빅 데이터 개념

구체적이고 정확한 정의는 없지만 전통적 개념은 구글이나 마이크로소프트 등 대기업이나 NASA의 연구 프로젝트에서 분석하는 방대한 양의 데이터를 말함. 그래서 빅데이터를 very large DB, extremely large DB, exterme Data, total data 등 다양한 용어로 부린다. 맥킨지 보고서에 따라 데이터베이스의 규모에 초점을 맞춘 정의는, 일반적인 DBMS로 저장, 관리, 분석할 수 있는 범위를 초과하는 대규모 데이터이다.

## 빅 데이터 속성

가트너의 애널리스트 더그 레이니(doug laney)는 연구 보고서에서 현재 가장 널리 사용하는 빅 데이터의 속성을 3V, 즉 **규모(volume)**, **다양성(variety)**, **속도(velocity)** 등 세가지로 정의하고 있으며, IBM에서는 **정확성(veracity)** 요소를 더하고, 최근에는 **가치(value)**를 포함하여 5V로 정의된다.


그림 11. 빅 데이터의 속성

### *규모(Volume)*

미디어나 위치 정보, 동여상 등과 같이 다루어야 할 데이터의 크기를 말하는 것이다. 물리적인 크기뿐만 아니라 현재의 기술로 처리 가능한 양인지, 불가능한 양인지에 따라 빅 데이터를 판단하며, 기술의 발달에 따라 킬로바이트, 메가바이트, 기가바이트, 최근에는 테라바이트를 훌쩍 넘어 요타바이트까지를 빅 데이터로 통칭한다.



### *다양성(Variety)*

다양성은 다양한 종류의 데이터를 수용하는 속성을 말하며, 빅 데이터는 형식이 정해져 있는 정형 데이터뿐만 아니라, 감시 카메라에서 생성되는 동영상, 개인이 디지털 카메라로 생성하여 웹 사이트에 올리는 사진, 소셜 네트워크 서비스로 전달되는 메시지, 물건에 부착되거나 주변에 설치된 센서에서 발생하는 RFID 태그나 센서 값 등 다양한 비정형 데이터도 포함된다.

데이터를 정형화 정도에 따라 **정형(structured)**, **반정형(semi-structured)**, **비정형(unstructured)**로 분류한다.

표2. 빅 데이터의 종류

| 종류 | 설명 |
|---|---|
| 정형 | 고정된 필드에 저장된 데이터<br>예) 관계형 데이터베이스, 스프레드시트[8] |
| 반정형 | 고정된 필드에 저장되어 있지는 않지만, 메타데이터나 스키마 등을 포함하는 데이터 예) XML, HTML 텍스트 |
| 비정형 | 고정된 필드에 저장되어 있지 않은 데이터<br>예) 텍스트 분석이 가능한 텍스트 문서, 이미지·동영상·음성 데이터 |

**정형 데이터:** 정형 데이터는 일정한 규칙에 따라 체계적으로 정리한 데이터이며, 정형화된 데이터는 그 자체로도 의미 해석이 가능하며, 바로 활용이 가능한 데이터를 포함한다.

**반정형 데이터:** 반정형 데이터는 한글이나 MS워드 등으로 작성한 데이터이며, 페이스북, 트위터, 카카오톡 등 소셜 네트워크 서비스 사용자가 생성하는 데이터들이 이에 해당된다.

**비정형 데이터:** 동영상, 음성, 센서, GPS, SNS 등에서 발생하는 데이터이다.

### *속도(Velocity)*

속도는 대용량의 데이터를 빠르게 처리하고 분석할 수 있는 속성을 말하며, 데이터를 자동으로 생성하는 센서, 스마트폰 등 데이터 생성 및 유통 채널의 다변화로 데이터 생성 속도가 빨라진다. 이는 처리 속도의 가속화를 요구한다.

### *정확성(Veracity)*

정확성은 데이터에 부여할 수 있는 신뢰 수준을 말하며, 높은 데이터 품질을 유

8) (spreadsheet)는 경리, 회계 등의 계산을 위해 사용되는 표현식의 계산용지나, 계산용지를 컴퓨터에서 사용할 수 있게 구현한 표 계산 프로그램을 의미함.

지하는 것은 빅 데이터의 중요한 요구 사항이자 어려운 과제이다. 하지만 최상의 데이터 정제(data cleansing)[9]기법을 사용해도 날씨나 경제, 고객의 미래 구매 결정 같은 일부 데이터의 본질적인 불 확신성은 제거할 수 없다.

소셜 네트워크 같은 인간 환경에서 생산되는 데이터는 신뢰하기가 어렵고, 미래는 예측하기 어려우며, 사람과 자연, 보이지 않는 시장의 힘 등이 빅 데이터의 다양한 불확실성 형태로 나타나다.

***가치(Value)***

가치는 빅 데이터를  저장하려고 IT 인프라 구조 시스템을 구현하는 비용을 말한다. 빅 데이터의 규모는 엄청나며 대부분은 비정형적인 텍스트와 이미지 등으로 구성되어 있다. 이 데이터들은 시간이 지남에 따라 빠르게 전파하면서 변하므로 그 전체를 파악하고 일정한 패턴을 발견하기가 쉽지 않아 가치의 중요성이 강조되고 있다.

그리고 광의의 빅 데이터 정의는 데이터 자체와 이를 분석하는 전문가 및 조직 그리고 데이터를 처리·축적·분석 기술 등을 포괄적으로 포함될 수 있다. 즉, 빅 데이터만으로는 아무런 가치를 획득할 수 없고 빅데이터 분석 전문가 그리고 처리 기술 등이 확보되어야만 원하는 가치를 얻을 수 있기 때문이다.

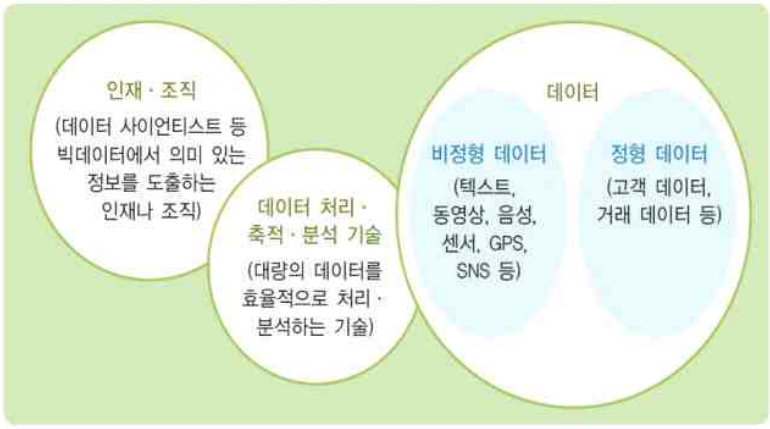



그림 12. 광의의 빅 데이터 정의

9) 정제. data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. 데이터 청소 또는 정제, 세척이라고도 불림, 데이터의 품질을 높이기 위해 데이터를 검출, 에러와 불일치를 제거함.



빅 데이터와 전통적 데이터는 차이는 비단 데이터의 크기뿐만 아니라 다양한 관점에 큰 차이를 보인다.

표3. 전통적 데이터와 빅 데이터와의 비교

| 구분 | 전통적 데이터 | 빅 데이터 |
|---|---|---|
| 데이터 원천 | 전통적 정보 서비스 | 일상화된 정보 서비스 |
| 목적 | 업무와 효율성 | 사회적 소통, 자기표현, 사회 기반 서비스 |
| 생성주체 | 정부 및 기업 등 조직 | 개인 및 시스템 |
| 데이터 유형 | ◦정형 데이터<br>◦조직 내부 데이터(고객 정보, 거래 정보 등)<br>◦주로 비공개 데이터 | ◦비정형 데이터(비디오 스트림, 이미지 오디오, 소셜 네트워크 등 사용자 데이터, 센서 데이터, 응용 프로그램 데이터 등)<br>◦조직 외부 데이터<br>◦일부 공개 데이터 |
| 데이터 특징 | ◦데이터 증가량 관리 가능<br>◦신뢰성 높은 핵심 데이터 | ◦기하급수로 양적 증가<br>◦쓰레기(Garbage) 데이터 비중 높음<br>◦문맥 정보 등 다양한 데이터 |
| 데이터 보유 | 정부, 기업 등 대부분 조직 | ◦인터넷 서비스 기업(구글, 아마존 등)<br>◦포털(네이버, 다음 등)<br>◦이동 통신 회사(SKT, KTF 등)<br>◦디바이스 생산 회사(애플, 삼성전자 등) |
| 데이터 플랫폼 | 정형 데이터를 생산·저장·분석·처리할 수 있는 전통적 플랫폼<br>예) 분산 DBMS, 다중처리기, 중앙집중 처리 | 비정형 대량 데이터를 생산·저장·분석·처리할 수 있는 새로운 플랫폼<br>예) 대용량 비정형 데이터 분산 병렬 처리 |

## 빅 데이터 처리 과정

아래 그림은 빅 데이터를 처리하는 과정을 크게 데이터의 **생성, 수집, 저장, 처리, 분석, 표현**의 과정으로 분류한 것이다.


그림 13. 빅 데이터 처리 과정

빅 데이터를 처리되는 특징은 아래 표와 같다.

표4. 빅 데이터의 처리 특징

| 구분 | 처리 특징 |
|---|---|
| 의사 결정 속도 | 빠른 의사 결정이 상대적으로 덜 요구되어 장기적·전략적 접근 필요 |
| 처리 복잡도 | 다양한 데이터 소스, 복잡한 로직 처리, 대용량 데이터 처리로 처리 복잡도가 높아 분산 처리 기술 필요 |
| 데이터 규모 | 처리할 데이터 규모가 방대. 즉, 고객 정보 수집 및 분석을 장기간에 걸쳐 수행해야 하므로 처리해야 할 데이터양이 방대 |
| 데이터 구조 | 비정형 데이터의 비중이 높음. 즉, 소셜 미디어 데이터, 로그 파일, 스트림 데이터, 콜센터 로그 등 비정형 데이터 파일의 비중이 높음 |
| 분석 유연성 | 처리·분석 유연성이 높음. 즉, 잘 정의된 데이터 모델, 상관관계, 절차 등이 없이 기존 데이터 처리 방법에 비해 처리 및 분석 유연성이 높음 |
| 처리량 | 동시 처리량이 낮음. 즉, 대용량 및 복잡한 처리가 가능하여 동시에 처리할 수 있는 데이터양이 적어 실시간 보장되어야 하는 데이터 분석에는 부적합 |



## 빅 데이터 처리 기술

빅 데이터 처리 과정에서 여러 가지 기술들이 등장하였음. 각 과정별 기술 영역을 정리하면 아래와 같다.

표5. 빅데이터 처리 과정별 기술 영역

| 과정 | 영역 | 개요 |
|---|---|---|
| 생성 | 내부 데이터 | 데이터 베이스, 파일 관리 스시템 |
| | 외부 데이터 | 인터넷으로 연결된 파일, 멀티미디어, 스트림 |
| 수집 | 크롤링(Crawing) | 검색 엔진의 로봇을 사용한 데이터 수집 |
| | ETL(Extraction Transformation Loading) | 소스 데이터의 추출·전송·변환·적재 |
| 저장 | NoSQL 데이터베이스 | 비정형 데이터 관리 |
| | 스토리지 | 빅데이터 저장 |
| | 서버 | 초경량 서버 |
| 처리 | 맵리듀스 | 데이터 추출 |
| | 프로세싱 | 다중 업무 처리 |
| 분석 | NLP(Neuro Lingustic Programing) | 자연어 처리 |
| | 기계 학습(Machine Learning) | 기계 학습으로 데이터의 패턴 발견 |
| | 직렬화(Serialization) | 데이터 간의 순서화 |
| 표현 | 가시화 | 데이터를 도표나 그래픽적으로 표현 |
| | 획득 | 데이터의 획득 및 재해석 |

### *빅 데이터 소스 생성과 수집 기술*

데이터는 소스 위치에 따라 내부 데이터와 외부 데이터로 구분함. 따라서 데이터 수집도 내부, 외부 수집으로 구분할 수 있으며, **내부 데이터 수집**은 주로 자체적으로 보유한 내부파일 시스템이나 데이터베이스 관리 시스템, 센서 등에 접근하여 정형 데이터를 수집한다.

**외부 데이터 수집**은 인터넷으로 연결된 외부에서 비정형 데이터를 수집하고, 데이터 수집은 주로 아래그림과 같은 툴, 프로그래밍으로 자동으로 진행된다.

표6. 빅데이터 자동 수집 방법

| 방법 | 설명 |
|---|---|
| 로그 수집기 | 내부에 있는 웹 서버의 로그를 수집. 즉, 웹로그, 트랜잭션 로그, 클릭 로그 데이터 등 수집 |
| 크롤링 | 주로 웹로봇으로 거미줄처럼 얽혀 있는 인터넷 링크를 따라다니며 방문한 웹 사이트의 웹 페이지라든가 소셜 데이터 등 인터넷에 공개된 데이터 수집 |
| 센싱 | 각종 센서로 데이터 수집 |
| RSS[10] 리더/오픈 API | 데이터의 생산공유참여 환경인 웹2.0을 구현하는 기술로 필요한 데이터를 프로그램밍으로 수집 |
| ETL[11] | 데이터의 추출, 변환, 적재의 약자로, 다양한 소스 데이터를 취합해 데이터를 추출하고 하나의 공통된 형식으로 변환하여 데이터웨어하우스[12]에 적재하는 과정 지원 |

### *빅 데이터 저장 기술*

데이터에서 의미 있는 정보를 추출하려면 효율적으로 저장 관리하는 기술이 필요하다. 데이터 저장 관리는 추후 사용할 수 있도록 데이터를 안전하고 효율적으로 저장하는 것으로 빅 데이터는 대용량, 비정형, 실시간성 속성을 수용할 수 있는 저장 방식이 필요하며, 특히 대량의 데이터를 파일 형태로 저장할 수 있는 기술과 비정형 데이터를 정형화된 데이터 형태로 저장하는 기술이 중요하다. 그 대표적인 기술은 아래 그림과 같다.

표7. 대용량 데이터를 저장하는 다양한 접근 방식

| 접근 방식 | 설명 | 제품 |
|---|---|---|
| 분산 파일 시스템 | 컴퓨터 네트워크로 공유하는 여러 호스트 컴퓨터 파일에 접근할 수 있는 파일 시스템 | GFS(Google File System), HDFS(Hadoop Distributed File System) 아마존 S3 파일 시스템 |
| NoSQL | 데이터 모델을 단순화해서 관계형 데이터 모델과 SQL을 사용하지 않는 모든 DBMS 또는 데이터 저장 장치 | Cloudata, HBase, Cassandra |
| 병렬 DBMS | 다수의 마이크로프로세서를 사용하여 여러 디스크의 질의, 갱신, 입출력 등 데이터베이스 처리를 동시에 수행하는 데이터베이스 시스템 | VoltDB, SAP HANA, Vertica, Greenplum, Netezza |
| 네트워크 구성 저장 시스템 | 서로 다른 종류의 데이터 저장 장치를 하나의 데이터 서버에 연결하여 총괄적으로 데이터를 저장 및 관리하는 시스템 | SAN(Storage Area Network) NAS(Network Attached Storage) |

10) RSS(rich site summary) 뉴스나 블로그 사이트에서 주로 사용하는 콘텐츠 표현 방식임.
11) ETL이란 데이터 웨어하우스(data warehouse; DW) 구축 시 데이터를 운영 시스템에서 추출하여 가공(변환, 정제)한 후 데이터 웨어하우스에 적재하는 모든 과정을 말함. ETL은 데이터 추출(extraction), 변환(transformation), 적재(loading)의 약자임.

12) 데이터 웨어하우스란 사용자의 의사 결정에 도움을 주기 위하여 기간시스템의 데이터베이스에 축적된 데이터를 공통의 형식으로 변환해서 관리하는 데이터베이스를 말함. 방대한 조직 내에서 분산 운영되는 각각의 데이터베이스 관리 시스템들을 효율적으로 통합하여 조정, 관리하여 효율적인 의사 결정 시스템을 위한 기초를 제공함.



### *빅 데이터 처리 기술*

빅 데이터는 방대한 양의 데이터와 데이터 생성 속도, 데이터 종류의 다양성을 통합적으로 고려할 수 있는 기술이 필요하다. 하둡은 빅 데이터 기술로는 정형, 비정형 빅 데이터 분석에 가장 선호되는 솔루션이다.

맵리듀스 기술은 일반 범용 서버로 구성된 군집화 시스템을 기반으로 <키, 값> 입력 데이터 분할 처리 및 처리 결과 통합 기술, Job 스케줄링 기술, 작업 분배 기술, 장애에 대처하는 태스크 재수행 기술 등이 통합된 분산 컴퓨팅 기술이며, R은 R언어와 개발 환경으로 기본적인 통계 기법부터 모델링, 최신 데이터 마이닝 기법까지 구현 및 개선이 가능하다. NoSQL은 전통적인 관계형 데이터베이스 RDBMS와는 다르게 설계된 비관계형 데이터베이스이다.

### *빅 데이터 분석 기술*

빅 데이터 분석에 사용하는 기술은 대부분 통계학과 전산학, 특히 기계 학습과 데이터 마이닝 분야에서 이미 사용한 것들이다. 이 분석 기술들의 알고리즘을 대규모 데이터 처리에 맞게 개선하여 빅 데이터 처리에 적용시키고 있는 것이다. 사용할 수 있는 대표적인 분석 기술은 아래 그림과 같다.

표8. 빅 데이터 분석 기술

| 용어 | 설명 |
|---|---|
| 텍스트 마이닝 | 자연어 처리 기술을 사용해 인간의 언어로 쓰인 비정형 텍스트에서 유용한 정보를 추출하거나 다른 데이터와의 연계성을 파악하여, 분류나 군집화 등 빅 데이터에 숨겨진 의미 있는 정보를 발견하는 것 |
| 웹 마이닝 | 인터넷에서 수집한 정보를 데이터 마이닝 기법으로 분석하는 것 |
| 오피이언 마이닝 | ◦다양한 온라인 뉴스와 소셜 미디어 코멘트, 사용자가 만든 콘텐츠에서 표현된 의견을 추출·분류·이해하고 자산화하는 컴퓨팅 기술<br>◦텍스트 속의 감성과 감동 여러 가지 감정 상태를 식별하려고 감성 분석 사용<br>◦마케팅에서는 버즈(Buzz; 입소문) 분석이라고도 함 |
| 리얼리티 마이닝 | ◦휴대폰 등 기기를 사용하여 인간관계와 행동 양태 등을 추론하는 것<br>◦통화량, 통화 위치, 통화 상태, 대상, 내용 등을 분석하여 사용자의 인간관계, 행동 특성 등 정보를 찾아냄 |
| 소셜 네트워크 분석 | 수학의 그래프 이론을 바탕으로 소셜 네트워크 서비스에서 소셜 네트워크 연결 구조와 연결 강도를 분석하여 사용자의 명성 및 영향력을 측정하는 것 |
| 분류 | ◦미리 알려진 클래스들로 구분되는 훈련 데이터군을 학습시켜 새로 추가되는 데이터가 속할 만한 데이터군을 찾는 지도 학습 방법<br>◦가장 대표적인 방법으로 KNN(K-Nearest Neighbor)이 있음 |
| 군집화 | ◦특성이 비슷한 데이터를 합쳐 군으로 분류하는 학습 방법<br>◦분류와 달리 훈련 데이터 군을 이용하지 않기 때문에 비지도 학습 방법<br>◦트위터에서 주로 사진/카메라를 논의하는 사용자군과 게임에 관심 있는 사용자군 등 관심사나 취미에 따라 분류 |

표8(계속). 빅 데이터 분석 기술

| 용어 | 설명 |
|---|---|
| 기계 학습 | ◦인공지능 분야에서 인간의 학습을 모델링한 것<br>◦컴퓨터가 학습할 수 있도록 하는 알고리듬과 기술을 개발하여 수신한 이메일의 스팸 여부를 판단할 수 있도록 훈련<br>◦결정 트리 등 기호적 학습. 신경망이나 유전자 알고리듬 등 비기호적 학습, 베이지안(Baysian)이나 은닉 마크프(Hidden Markov) 등 확률적 학습 등 다양한 기법이 있음 |
| 감성 분석 | 문장의 의미를 파악하여 글의 내용에 긍정/부정. 좋음/나쁨을 분류하거나 만족/불만족 강도를 지수화, 그런 다음 이 지수를 이용하여 고객의 감성 트랜드를 시계열적으로 분석하고 고객 감성 변화에 기업의 신속한 대응 및 부정적인 의견의 확산을 방지하는 데 활용 |

***빅 데이터 표현 기술***

데이터 분석 결과를 효과적으로 전달하려고 어렵고 복잡한 정보를 한눈에 쉽게 이해할 수 있도록 간단한 도표나 3D 이미지 등으로 표현하는 정보 표현 기술이 발전하였다. 2009년 구글에서 개발한 Fusion Tables은 방대한 양의 데이터를 표현해 주는 온라인 서비스이다. 다음의 그림은 정보 표현의 간단한 예이다.

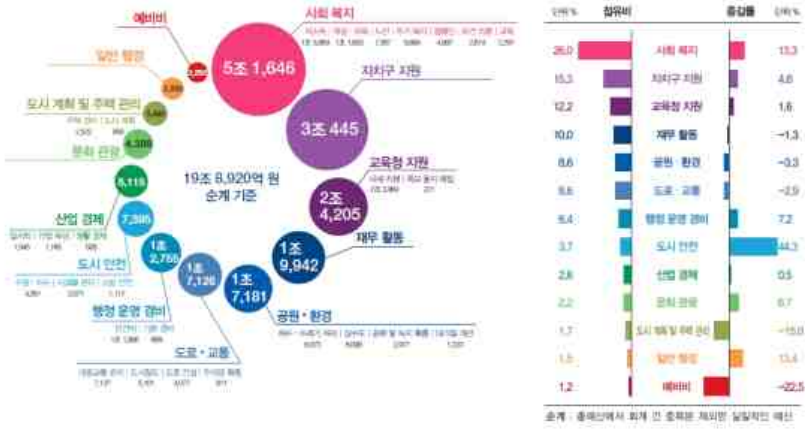



그림 14. 정보 표현의 간단한 예

# 빅 데이터 적용 사례

예측 기반 제조 제어(Predictive Manufacturing Control)에 대한 빅 데이터 분석 활용: 제조 공정 산업체에 대하여

요즈음 기업들은 국제적인 경쟁 시장의 극심하게 증가하고 있는 요구사항을 충족시키기 위하여 기존에 존재하는 사업 환경을 새로운 사업 실행 과정 시스템으로 개선하여 적용하고 있는 추세이다. 제품 생산 공정에 대하여 사전에 불량품을 예측하고 최적의 생산 공정에 대한 설계를 빅 데이터 활용 기반의 예측 기반 제조 제어 기법에서 해법을 찾고 있다. 특히, 독일 최대 철강 제품 생산 기업인 Saarstahl AG는 이러한 빅 데이터 기술을 활용하여 획기적인 생산 효율성을 확보하고 있다.

사물인터넷(Internet Of Things; IOT) 시대의 도래를 통해 현 시대에서는 대부분의 데이터가 디지털화되고 있다. 이러한 변화는 제조 산업체에서도 적용되어 제조 공정의 생산력을 증대하는 역할로 활용되고 있는 추세이다. 제조 산업체에서의 빅 데이터 활용의 예로 업무 과정 관리(Business Process Management; BPM)를 들 수 있는데, 이를 통해 기업은 작업 과정의 실행 분석, 향후 이어지는 과정의 예측 그리고 발생 가능한 문제를 파악할 수 있게 된다. 이는 기업 내에 모든 작업 과정에서 발생하는 데이터를 수집, 저장, 그리고 분석 등을 통해 가능하게 된다. 일반적으로 기업 내의 작업 과정은 매우 복잡한 사건들(Events)이 서로 연계되어 진행되고 있다. 현재의 산업체에서의 빅데이터에 대한 접근과 기술은 이러한 복잡하고 연속적인 사건 과정(Complex Event Processing; CEP)을 어떠한 처리하느냐에 집중되고 있는 상황이다.

사건 기반 예측(Event-based Predictions) 기법은 사전 계획에 대한 '다양한 후보군 색출', '향후 발생 가능한 오류나 제품 품질에 대한 예측' 그리고 '작업 과정의 제어' 등에 활용될 수 있다. 이러한 정보 시스템을 구축하기 위해서 기업은 기본적은 질문에 답을 가지고 있어야 한다: 첫째, 사건 기반 예측 시스템을 구현하기 위한 데이터 감지 기술(Sensor Technology)을 이용하여 획득하게 되는 데이터의 유형이 무엇인지와, 둘째, 이러한 데이터 분석하는 목적이 무엇인지 등을 명확히 해야만 한다.

### *데이터 분석(Analytical Process) 기반의 제조 산업에서의 제품 생산 공정*

일반적으로 제조 산업은 크게 네가지 유형으로 분류되는데, 입력 요소와 이와 연관된 산출물과의 관계를 통해 정의되고 있다. "Continuous 제조 과정", "Analytical 제조 과정", "Synthetical 제조 과정", 그리고 "Mixed analytical & synthetical 제조 과정" 등으로 분류 된다(그림 15 참조).



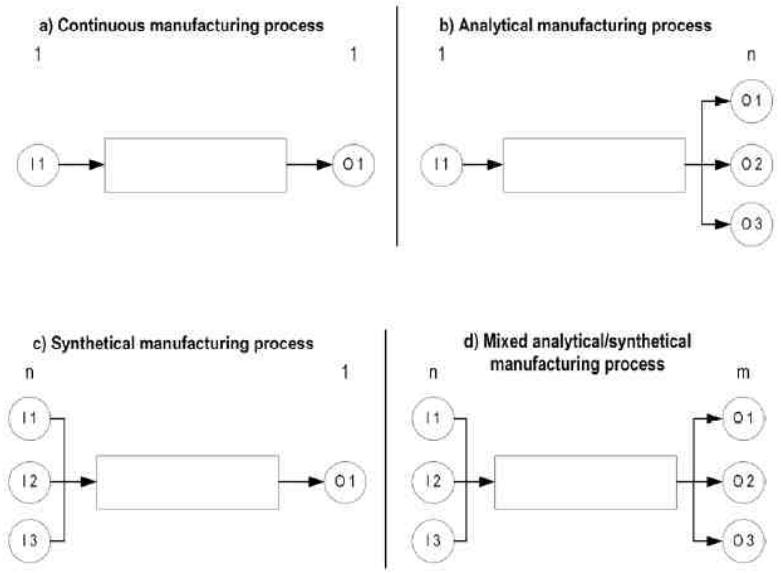



그림 15. 제조 과정에서의 존재한 실제 제조 형태에 대한 개념

이러한 통상적인 제조 산업체에서의 제조 공정과 데이터 분석 기반 제조 산업은 주된 제품(main productions)과 함께 생산되는 제품(co-productions) 그리고 부산물(by-productions) 등을 생산할 때, 모든 제품 생산 공정에 대한 생산 예측과 조정을 상시적으로 진행하여 역동적으로 데이터를 활용한다. 이러한 데이터 분석을 통해 생산 제품의 품질 변화를 사전에 감지하게 되고 즉각적으로 오류를 수정할 수 있게 된다. 그러나 아직까지 정밀하고 즉각적인 예측과 조정 등을 위한 빅데이터 처리 기술은 매우 복잡하고 어려운 난재이다.

### *사건 기반 예측 기법을 활용하는 데이터 분석 제조 과정에서의 예측과 조정*

제조 기업은 제품 생산에 대한 각종 요구 조건, 즉, 매우 복잡한 계산을 통해 산출되는 제료비 등을 관리하기 위하여 ERP(Enterprise Resource Planning) 기법을 채용하게 된다. 이러한 기법은 제조 과정에 대한 예측 기반 제어가 거의 불가능하게 된다. 산출된 예측에 활용되는 데이터와 양이 너무나 부정확하고 소량인 경우가 대부분이며 제조 공정 상황을 파악하기 위해 필요한 데이터가 턱없이 부족하다. 그리고 ERP 시스템에서의 데이터 확충은 너무 복잡하고 고비용하기 때문에 거의 실현 불가능한 상황이다. 이런 시스템을 활용하여 고품질의 제품을 생산을 위한 사전 예

측과 실시간 조정과는 거리가 먼 미리 정해놓은 절차대로 생산하게 되는 것이다. 즉, 이러한 ERP 시스템은 획득된 데이터를 기반으로 불충분한 분석과 예측만을 제공하게 되는 것이다.

"사물인터넷(IOT)"와 "가상 물리 시스템(Cyber-Physical Systems)"에 연구가 최근 활발히 진행되고 있다. 이러한 새로운 환경은 상대적으로 저가로 제품 생산 공정에서 파생되는 데이터를 축적하고 있게 하고 실시간으로 생산 공정에서의 내외부 생산 인자들을 측정할 수도 있게 한다. 이러한 시스템은 예측 과정에서 대단히 중요한 정보로 활용되는 데이터의 지속적인 축적을 통해 완성된다.

사건 기반 제품 생산 예측과 제어는 기본적으로 크게 네 가지 요소를 고려되어야 한다. 첫째, 제조 과정에서 나오는 각종 인자들의 데이터 유형에 대한 파악, 둘째, 복잡한 사건의 패턴에 대한 분석, 셋째, 전통적인 제조 공정 데이터와 상황 인지 정보(Contextual Information)의 연관성 파악(이는 사건 기반 공정 예측과 제어에 핵심적 역할을 하게 된다), 넷째, 제품 생산 공정 예측 등이다. 이러한 복잡하지만 정교한 계산 과정을 통하여 제조 기업은 보다 상위의 제품 생산에 대한 예측과 제어를 할 수 있게 되는 것이다(그림 16 참조).

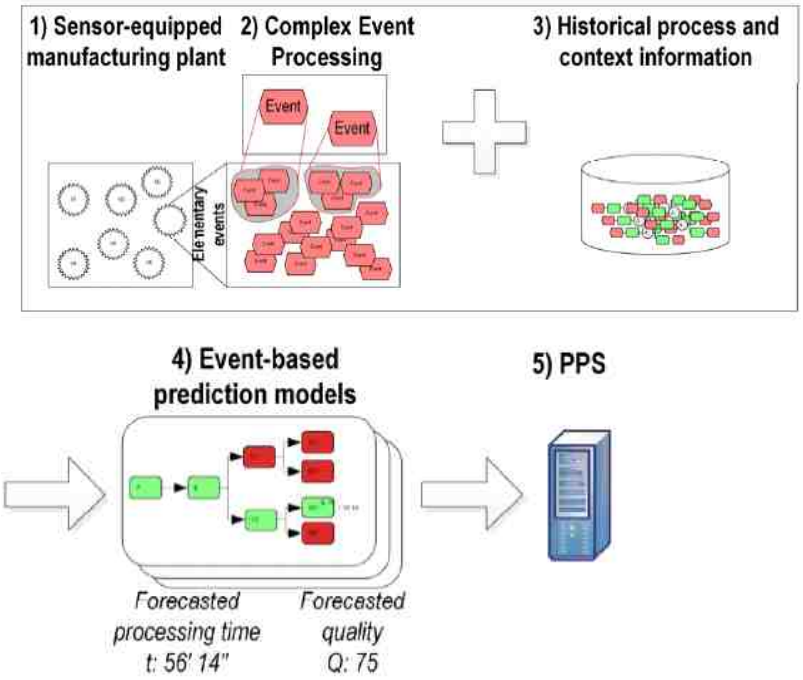



그림 16. 사건 기반 제품 생산에 대한 예측과 제어의 기본 개념도



## 제품 개발(Product Development) 과정에서의 빅 데이터

성공적인 신제품 개발은 기업 제품 출시가 시장에서의 경쟁력을 확보하기 위한 핵심적 요소이다. 요즈음과 같은 불경기에는 크고 작은 모든 기업이 신제품 개발을 효율적으로 하는 것은 매우 중요시 되고 있다. 산업 정보와 지식을 적극 활용하여 신속하고 성공적인 신제품 생산(New Product Development; NPD)과정 구축에 많은 기업이 집중하고 있는 추세이다.

기업들은 신제품 출시(New Product Induction; NPI)에 있어서 지식 관리(Knowledge Management)의 중요성을 인식하고 있어 왔다. 특히, NPI 과정에는 다양한 단계에서의 입력 요소로서의 데이터, 정보 그리고 지식 등의 조합이 활용된다. 최근 몇 년 동안, 기업들은 다양한 소스로부터 발생되는 데이터 활용에 대한 필요성을 깨달기 시작하였다. 그리고 이러한 데이터는 제품 성능의 개선에 막강한 능력을 제공하기 때문에 기업은 NPI 과정에 데이터를 활용하여 제품 생산 과정의 효율화, 비용 절감 그리고 소비자 만족도 향상 등에 많은 효과를 거두고 있다.

이러한 데이터로부터의 효과를 거두기 위해서, 기업은 다양한 사업 기능에서 요구되는 지식의 형태로 데이터를 수집, 분석 그리고 전파 등을 위하여 잘 정리된 활용 전략이 반드시 필요하다. 예를 들면, 제품 아이디어, 소비자 행동 패턴, 고객의 목소리, 제품 품질 기능(Quality Function Deployment) 데이터 그리고 소셜 네트워크에서의 제품 경향 등은 제품 생산에 대한 전략과 포크폴리오 설계에 큰 도움을 줄 수 있을 것이다. 또한, 품질 보증, 품질 평가 데이터, CAD 시스템에서 파생된 데이터 그리고 제조 과정에서의 데이터 등도 역시 NPD 과정으로 피드백되는 정보로서 제품 설계 및 검증 등에 도움을 줄 수 있다.

제품 개발에 대한 조직적 지식 데이터는 단순한 형태로 활용되는 것은 아니며, 일반적으로 방대한 양이고 기업들 사이에 널리 퍼져있다. 제조 산업에서의 제품 개발에 대한 NPD 과정에서의 빅 데이터 분석을 위한 지식 데이터 관리에 대하여 살펴보겠다.

### *제조 기업의 당면 과제 극복*

새로운 경제의 급부상, 지역 및 세계 시장에서 제품 공급에 대한 증가는 요구사항, 그리고 제품 수명 주기에 걸쳐있는 정보 기술의 중요성의 증가 등은 제품의 상용화를 이끌고 있다. 소비자는 '왕'이며 기업의 흥망을 결정한다. 뿐만 아니라 다양한 시장에 출시되는 맞춤형 제품에 대한 포트폴리오를 확보하기 위하여, 기업은 서로 다른 영역에서의 소비자 요구에 대한 확실한 이해가 필요하며 소비자의 기대를 충족시킬 수 있는 제품을 설계해야만 한다. 글로벌 시장에서의 유사한 제품에 대한 첨예한 경쟁은 이러한 복잡성을 가중시키고 있다.

이러한 추세에서 '시기 적절한 출시(time to market)', '첫 시장 출시(first time right)', 그리고 '가격 경쟁력(cost competitiveness)' 등이 높은 경쟁 산업체들(자동차

생산 업체, 소비제 생산 업체, 그리고 가전 제품 생산 업체 등)에 진행되는 OEM 방식의 제조 업체들 사이에서 중요한 차별화 중요로 부각되고 있다. 오늘날의 경쟁은 신 제품을 생산하는 것이 아니라 오히려 얼마나 빠르게 생산하느냐 하는 것이다. 그러므로 기업들은 신 제품을 신속하게 생산할 수 있게 하는 상시적인 지식 저장 체계를 구축하기 위하여 현재와 과거 시장 정보, 제품 디자인 정보, 제조 과정에서 파생되는 데이터 그리고 제품 성능 시험 및 서비스 데이터 등의 활용에 지대한 관심을 가져야만 한다.

### *데이터, 정보 그리고 지식*

현 시대의 지식 체계(그림 ?? 참조)에서 데이터, 정보, 지식 그리고 지혜 등을 구분하는 것은 매우 중요하다. 데이터는 사건들과 관련해서 발생하는 개별적이고 목적 지향의 사실들의 집합이다. 정보는 문서 혹은 음성-화상 기반의 의사소통 형태의 메시지이다. 어떠한 메시지에서도 공급자와 수혜자가 존재한다. 정보는 수혜자 무언가를 인지하는 방법을 바꿀 수 있다는 것과 수혜자의 판단과 그 후의 행동에 영향을 줄 수 있다는 것을 의미한다. 지식은 데이터와 정보 보다 훨씬 더 넓고 깊으며 풍부하다. 지식은 데이터 혹은 정보 등의 응용, 분석 그리고 생산적 사용으로부터 파생한다.

제품 개발 과정의 핵심은 지식과 이의 재사용 속에 존재한다. 지식은, 효과적으로 관리될 때, NPI 과정에 있어서의 비용 절감, 제품 품질 향상 그리고 소비자 만족도 증가 등이 가능해진다. 지식 기반 기업에서 지식은 기업 행동을 안내하고 지속 가능한 경쟁적 장점을 확립하는 데에 결정적 역할을 하게 된다. 전형적으로 신 제품 생산 과정에서는 데이터, 정보 그리고 지식 등을 조합하여 이용된다. 지식 관리 피라미드에서의 지혜는 기업 내에 있는 전문가와 몇몇 안되는 개인으로 부터 출현되는 것으로 통상적으로 우수 사례와 학습된 성과물 등으로 나타난다.

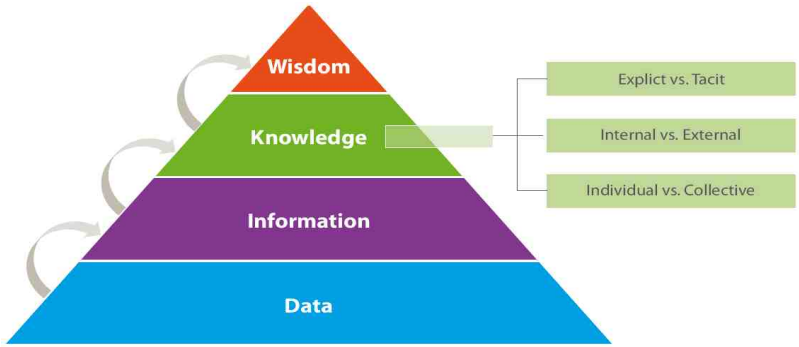



그림 17. 지식 관리 피라미드



### *제품 개발을 위한 집합체로서의 지식 관리*

전통적인 제품 개발 과정(그림 18 참조)은 생산 계획에서부터 제품 설계, 검증 그리고 실험에 까지 포함하며 최종적으로 제품 사용 주기 이후 서비스 계획으로 끝이 난다.


그림 18. 제품 개발 절차

제조 기업에서의 제품 개발 과정에서 몇몇의 데이터 형태가 종종 진행되어 재사용 가능한 지식 기반의 거대 분량의 데이터로 생성되어, 실시간 지능을 제품 생산팀에게 제공하게 된다. 몇몇 경우에 있어서, 소비자 직관, 경쟁적 지능 그리고 제품 성능 데이터 등과 같은 종류의 데이터는 제품의 글로벌 경쟁력을 증가시키는 데에 큰 가치를 제공한다.

지식 관리 상에서 많은 연구가 가능하겠지만, 위에서 언급한 데이터의 요소로부터 지식의 통합과 이에 따른 제품 개발 과정에 이르기까지의 연구는 미흡한 상황이다. 그러므로 다양한 데이터 원천으로부터 유래된 지식의 관리와 제품 개발 과정에서의 이러한 데이터의 활용은 제품 개발 과정의 효율화에 대한 막강한 역량을 제공할 수 있을 것이다.

### *제품 개발 과정에서의 빅 데이터 활용*

하나의 기업에서는 통상적인 기반에서도 막대한 양의 데이터를 양산하고 저장한다. 이는 제품 설계, 분석, 시험 그리고 제품에 대한 서비스 등의 결과로부터 얻어진다. 이는 종종 '빅 데이터'로 불리어진다. 이러한 데이터가 어떻게 지식으로 변환되는 지에 대하여 살펴본다.

빅 데이터는 일반적으로 세 가지 차원으로 정의되는데, 즉, 규모(Volume), 속도(Velocity), 그리고 다양성(Variety)이다. 최근에 Oracle은 여기에 하나, 가치(Value)를 보태어 정의하고 있다:

- o 규모: 기계에서 파생되는 데이터이다.
- o 속도: 이는 데이터 처리 속도를 의미하거나 데이터 분석에 적용되는 시간 지체 간격이 줄어드는 개념으로도 이해된다. 그리고 원천 데이터를 가공하여 공정 과정에 적용하는 시간 개념까지 포함된다. 기계에서 생성되는 막대한 양의 데이터 만큼은 아니지만, 대중 매체의 데이터는 소비자 관계 관리에 유용한 의견과 관계 정립에 커다란 입력 정보를 제공하기도 한다.
- o 다양성: 데이터 입력(예, 소비자 직관, 경쟁적 지능, 소비자 경향, 벤치마킹 데이

터, 표준화, 재료 물성치 등)의 다양성을 의미한다. 이러한 데이터는 CAD/CAM/CAE 데이터, 도면, 문서, 시험 데이터, 제품 불량 데이터, 생산 과정 정보 데이터 등과 같은 결과물로 생성된다.

o 가치: 서로 다른 데이터의 경제적 가치는 데이터 원천과 최종 활용처에 상당히 연관된다.

수십 년 동안 전자 제품과 IT 기술의 활발한 활용은 기업이 직간접적으로 혹은 과거와 현재에서 방대한 양의 데이터를 처리할 수 있게 하였다. 제품 설계, 분석, 공학해석, 시험, 그리고 설계 검증 과정에 활용되는 대부분의 IT 시스템은 방대하고 다양한 종류의 데이터를 양산하고 있다. 제조 산업에서의 컴퓨터와 인베디드 시스템의 폭발적인 증가는 막대한 양의 비정형 그리고 정형 데이터 생성을 가속화하였다.

대부분의 기업들의 새로운 도전은 가용한 데이터로부터 의미있는 통찰력을 도출하여 이를 다시 지능적으로 적용하는 것이다. 지식 관리는 이러한 데이터를 효율적으로 관리하거나 제품 개발 과정에 이를 활용하는 최종 사용자에게 제공된다. 의미있는 정보와 통찰력을 발굴하고, 이를 다시 지식 데이스에 축적하며 최종 사용자에게 전달하기 위해 저장 관리하는 목적으로, 기업의 빅 데이터에는 직간접적인 데이터 원천으로부터의 데이터 수집, 상시적인 기업 데이터와 함께 분석 가공 등을 포함한다.

빅 데이터에서 최대의 가치를 도출하기 위해서, 기업들은 반드시 시간적으로 변화하는 데이터의 형태에 대하여 거대 분량의 데이터 축출과 전달 등을 신속히 할 수 있는 IT 인프라를 개선해야만 한다. 이는 기업내 데이터와 데이터 분석을 통합해야함을 의미한다. 사실상, 구형 시스템을 운용하는 기업들은 과거 데이터에 대한 분석과 아울러 현존하는 IT 시스템으로 시스템 개선을 고려해야만 하는 것이다.

Yuan, Yoon, and Helendar의 연구에 따르면, 지식 영역은 '시장 지식(Market Knowledge)', '인적 지식(Human Knowledge)', '기술 지식(Technology Knowledge)', 그리고 '과정 지식(Procedural Knowledge)' 등의 4 가지 형태로 분류될 수 있다(표 ?? 참조)

**생산 데이터 활용**

대부분의 제조 업체는 CAD, 공학해석, 제조 생산 그리고 제품 개발 관리 툴 등을 통해 발생된 제품 관련 데이터를 관리하는 IT 시스템을 확보하고 있다. 그러나 이러한 시스템에서 생성된 대규모 데이터는 시스템 내에서만 묶여 있는 상황이 대부분이다. 만약 이러한 데이터가 다른 시스템과 연계될 수만 있다면, 생산자가 새로운 가치를 얻을 수 있는 매우 의미있는 빅 데이터의 기회를 창출할 수 있다. 예를 들면, PLM[13]은 내외부의 이해 당사자들 사이에서의 제품 설계 협력을 위한 플랫폼을 제공할 수 있을 것이다. OEM 주문자는 공급자와 협력할 수 있고 신속하고 보다 향

13) PLM=Product Life Management



표 9. 빅 데이터 차원 기반의 제품 개발 과정에서의 지식 영역

| 지식 유형 | 규모 | 속도 | 다양성 | 가치 |
|---|---|---|---|---|
| 시장 지식 | ·고객 데이터<br>·경쟁자 데이터 | ·직접적 상호작용<br>·소셜 매체<br>·설문 | ·사장 분석<br>·통계 데이터<br>·벤치마킹 데이터<br>·경향 | ·고부가가치<br>·고객 데이터<br>·경쟁자 데이터 |
| 인적 지식 | ·경험 기반<br>·협력 | ·실시간 결정 | ·기술 기반<br>·경험 기반<br>·암묵 지식 | ·체험적인 |
| 기술 지식 | ·표준<br>·활용<br>·물성정보<br>·분야 데이터 | ·안전<br>·실시간 습득 | ·비용<br>·신뢰성<br>·패키지<br>·인체공학 | ·패턴 |
| 과정 지식 | ·설계 지식<br>·분석<br>·물성치 표준<br>·검증 시험 확인 지식 | ·설계 지식<br>·표준 물성치 | ·설계<br>·CAD/CAM/CAE<br>·분석<br>·제조 | ·CAD/CAM/CAE<br>·우수 사례<br>·시험/확인<br>·서비스 데이터<br>·제조 과정 데이터 |

상된 품질의 제품 개발을 지원하는 공급자의 기술과 지식에 영향력을 행사하기도 한다. 빅 데이터 분석을 통해 얻어진 유용한 지식과 원칙, 논리 등과의 연결 방법은 신속하고 적절한 초기 의사 결정에 도움을 주며, 제품 개발 시간과 비용을 획기적으로 줄이고 단축하게 하고, 데이터 재활용성 또한 개선시킨다.

제품 데이터는 특정 부품이 어떻게 설계되었는지와 어떤 도전적 항목에 봉착되었는지에 대한 정보를 제공한다. 활용 가능한 형태로 이러한 정보를 축출하는 것은 데이터 재활용을 가능하게 하는 이러한 지식을 형성하는 과정에서 첫 번째 단계이다. 두 번째 단계는 특정 프로젝트의 실행 중에 제품 설계자와 기획자 등에 의해서 사용될 수 있는 경영 혹은 기술적 논리/원칙 등의 형태로 이러한 지식을 변환하거나 제시하는 것이다. 예를 들면, Configurator[14]를 사용하는 기업체 이런 컴퓨터 프로그램이 새로운 부품과 조립체에 대한 설계 과정에서의 도움뿐만 아니라 기존 데이터베이스에서부터 날은 부품들을 거둬들이는 것으로의 표준화를 도모할 수 있게 한다. 제품 개발 가치 사슬로부터 얻어지는 데이터는 이러한 프로젝트의 실행을 가능하게 한다.

***고객의 의견***

기업은 신제품 개발에서 제품에 대한 요구 조건을 모으는 과정에서 고객의 의견이 반영되어야함을 인지하고 있다. 그러나 많은 제조 기업 종사자들은 기존 제품

14) Configurator: 옵션을 선택하고 제품이나 공정 결과 상의 변화를 표시할 수 있게 하는 컴퓨터 프로그램(신조어)

디자인들을 개선하거나 신 모델에 대한 사용 설명서 개발에 도움을 주는 목적으로 활용하기 위한, 늘어나는 고객의 의견 데이터로부터 결정적인 통찰의 주는 의견들을 아직까지도 체계적으로 획득하고 있지 못하고 있다. 최우량 제조 기업들은 제품 생산 시스템에 대한 부가적인 피드백로서 제품 보증 클레임, 품질 시험과 진단 등으로부터 발생되는 데이터를 수집할 수 있다. 이런 제조 기업들은 보다 나은 제품을 개발하기 위하여 피드백 근원들의 상호 연관 분석에도 많은 노력을 기우리고 있다. 제조 기업들이 획득하려는 새로운 데이터 원천에는 실질적인 제품 사용으로 기록된 소셜 매체 플랫폼과 센서 등의 기반에서 고객의 코멘트를 포함하고 있다. 보다 좋은 제품을 개발하고 고객과 관련된 복잡한 데이터를 얻고, 종합적인 영감을 얻기 위하여, 제조업체들, 제품 공급자들, 소매업자들, 그리고 다른 관련자들은 반드시 협력해야 하며 각자의 데이터를 서로 통합 활용해야 한다.

예를 들면, 이러한 데이터를 적절한 분석을 통해서, 한 통신 기기 제조업체는 2년 사이에 30% 정도의 총 매출액의 증대를 달성하였고, 불필요한 비용 지출을 제거하였으며 고객과의 유대 관계 개선의 효과도 부가적으로 얻고 있다. 그리고 고객들은 보다 높은 가격으로도 제품 구매를 하는 의지를 갖게 되었다.

### *내외부 이해 관계자들*

밀접한 이해 관계자들이 손 쉽고 시시적절한 활용이 가능한 빅 데이터의 구축은 엄청한 가치를 창출할 수 있다. 기업들은 연구개발 기능의 수행을 향상하기 위하여 빅 데이터 분석을 이용할 수 있다. 기업에서의 연구개발, 공학해석, 그리고 생산 제조 등으로부터의 얻어지는 데이터의 통합은 다시 설계해야만 하는 것으로 유발되는 비용 낭비를 크게 줄여주어 결국에는 제품의 신속한 시장 출시를 가속화해줄 수 있게 하는 공시공학 기술을 가능하게 한다.

혁신을 촉진시키고 새로운 고객의 요구에 부응하는 제품을 개발하기 위하여, 제조 기업들은 혁신적 채널을 통해 획득된 요소들에 많은 의존을 하고 있다. 몇몇 제조 기업들에서는, 웹 기반 플랫폼을 통해, 제품 개발에 대한 혁신적 사안 혹은 심지어 협력 등에 대한 아이디어를 제공받기 위한 목적으로 외부 이해 관계자들을 적극 활용하고 있다. Kraft와 P&G와 같은 소비제 제품 제조 기업들은 그들의 고객들에게서 의견을 청취하고 있으며 신제품을 개발하기 위하여 화학 분야 연구기관, 대학 및 산업체 연구개발자 등을 포함한 외부 전문가와 협력을 추진하고 있기도 한다.

그러나 이러한 오픈 혁신 전략 만큼이나 중요한 또 하나의 문제가 있다. 그것은 이해 관계자들 사이에서 발생되는 엉첨난 양의 입력 요소에서 어떻게 진정한 값진 가치를 축출하느냐이다. 빅 데이터 분석에 대한 도전적 문제는 제품 개발 과정에서 활용될 수 있는 적절한 정보를 축출하여 지능적인 분석을 가능하게 하는 기술 및 알고리듬에 대한 개발이다.

### *빅 데이터의 활성화*



수십 년 동안, 기업들은 관계적 데이터베이스에 매출 현황을 저장하여 경영적인 결정을 해오고 있어왔다. 그러나 이를 넘는 결정적 데이터로 인식되는, 비 전통적이며 잘 조직화되지 못하는 데이터 - 웹 블로그, 소셜 매체, 이메일, 센서, 그리고 사진들 -은 새로운 보물로 그 가치를 인정받고 있다. 저장 장치와 컴퓨팅 장비 등의 가격 하낙은 이러한 데이터 축적을 가능하게 하고 있다. 그 결과, 기업들은 점차적으로 비 전통적인 데이터 뿐만 아니라 가능성 높은 가치의 데이터, 즉, 빅 데이터에 많은 관심을 갖기 시작하고 있다.

제품 개발에 있어서 빅 데이터와 분석의 활용은 기업내에서의 실현 가능한 기술과 하드웨어와 소트프웨어 간의 밀접한 통합 등에 의하여 추진된다. 빅 데이터 전략은 향후 활용을 위해 저장 장치에 포관될 수 있는 지식을 생성하는 데이터를 센싱하고 획득하며 전송하여 변환, 저장 그리고 분석하는 기능을 포괄적으로 고려해야 한다.

그림 ??에서는 빅 데이터를 분석하는 우수한 접근법과 연속적인 지식 관리 체계에 대하여 보여주고 있다. 전형적인 제품 개발 과정은 ERP, CRM, PLM 등과 같은 기업내 산재되어 있는 시스템들과 상호 연계되어 있으며, 이를 통해 미래에 활용될 수 있는 지식을 축출하는 데에도 쓰이게 된다. 이러한 정보들은 명확하고 구조적이다. 정보와 데이터는 제품이 현실화되면서 관련된 시스템과 함께 연속적으로 바뀌게 된다. 비전통적이고 비정형 정보는 시뮬레이션, 센싱, 블로그, 고객 경험 등과 같은 곳에서 발생하고, 이것들은 향후에 활용되어야만 한다. 이러한 정보의 작은 부분이 전체 기업 시스템에 다시 적용되는 동안에 중앙 데이터 저장 장치에 다시 저장되어야만 한다.

제품 개발은 전형적으로 Stage-Gate 프로세스[15]이며 종종 고객, 서비스, 공급자, 판매 후 서비스 등으로부터의 각종 입력 요소들을 활용하는 경영적 기능과 통합된다. 제품 개발에서의 빅 데이터 활용은 제품 설계, 생산 제조, 품질, 보증 유효기간 등과 연관되어 발생되는 데이터를 통합하는 새로운 방법을 제공한다. 지식 저장 창고의 한 부분이 되는 정보를 끄집어내는, 데이터를 보관하고 동일한 시스템에서 분석을 수행하는 시스템의 활용이 나날이 증가할 것으로 예상한다.

### 제품 개발 과정에 빅 데이터 적용 전략

지식 관리 속에서 제조 산업체에서의 빅 데이터 활용이 제품 개발 과정을 획기적으로 개선시킬 수 있음을 살펴보았다. 보다 상위 수준의 제조 기업이 빅 데이터 활용을 위해서는 다음과 같은 전략이 필요할 것이다:

15) Stage-Gate : 성공적인 제품 개발을 목표로 아이디어 발의부터 출시까지의 제품 개발 전 과정을 관리하는 R&D로서 stage는 R&D 활동이 수행되는 단계를 의미하며, gate는 각 단계별 R&D 활동을 평가하고 중지, 계속 등 의사결정을 하는 관문을 의미한다

o 지식 관리에 대한 분명한 비전의 수립
o 감추고 싶은 영역, 미래 도전 과제 등을 포함하여 현재 상황에 대한 철저한 평가
o 지식 관리에 대한 장기적인 실천 전략 수립
o 비전과 전략의 성공적인 실행을 위한 로드맵 작성
o 경영 결정 과정, 기술, 인프라 그리고 조직 변화 관리 등에 집중된 경영 변화 프로그램으로서 지식 관리 실행

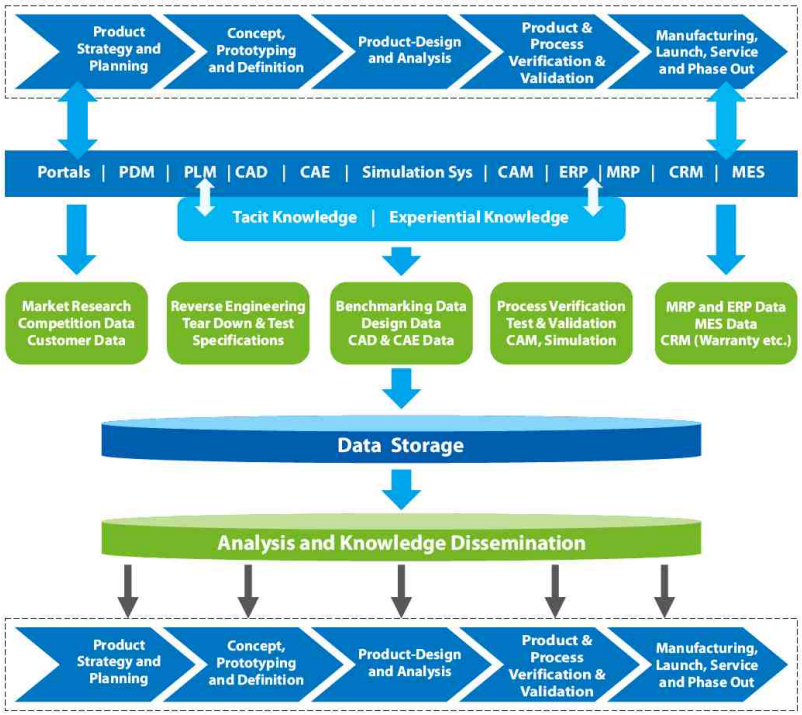


그림 19. 산업체 빅 데이터 원천과 지식 관리 체계도



## 빅 데이터 분석에 의한 선수육성실증실험: 가시아 레이솔 아카데미 사례

### *라이프 로그의 빅 데이터 분석에 의한 생활에서의 부가가치 창출*

사람의 행동을 24시간 센서로 기록·분석·가시화함으로써 생활이나 비즈니스에 가치를 파악하고, 새로운 부가가치의 창출을 가능하게 한다. 라이프 로그의 빅 데이터 분석은 개인의 건강관리뿐만 아니라, 경영적 관점에서 직원의 환경 개선까지 다양한 가능성을 내포하고 있는 분야이다. 히타치는 2012년 7월과 9월 가시와 레이솔 아카데미의 U-18의 선수를 대상으로 라이프 로그의 실증 실험을 실시하였다.

### *J 리그 진출을 목표로 하는 U-18 육성에 빅 데이터 활용*

가시와 레이솔의 선수 육성학교, 가시와 레이솔 아카데미

가시와 레이솔는 J2에서 승격하여 2011년 J리그 우승, 2013년 1월 제92회 천황배 전 일본 축구 선수권대회 우승으로 일본뿐만 아니라 세계로부터도 주목을 받고 있다. 가시와 레이솔 아카데미는 U-18, U-15, U-12, 스쿨 클래스는 연령별로 운영되는 축구 스쿨로 가시와 레이솔의 핵심 선수를 육성하는 조직이다. 기술 습득뿐만 나이라 인성 형성을 중점을 둔 선수 육성은 수많은 J리거 선수들을 배출하였으며, 2011년의 J리그 우승을 확정지은 배경에는  5명의 아카데미 출신 선수가 있다.


<일본 J 리그>

선수 육성에 빅 데이터 활용

가시와 레이솔 아카데미의 실증 실험은 센서 기술을 사용한 빅 데이터 분석에 가능성을 검증하는 것으로, 가시와 레이솔 아카데미의 선수의 협력 하에 퍼포먼스와 생활의 로그를 분석하였다.

### *데이터 분석 전문가에 의한 성공 프로세스*

프로젝트의 핵심은 기획력, 도메인 전문지식, 추진력

빅 데이터 프로젝트을 성공적으로 수행하기 위해서는 프로젝트의 의의와 가치 목적을 항상 명확히 정의하는 "기획력", 업무에 관한 깊은 지식과 빅 데이터 기술에 대한 "도메인 지식", 그리고 프로젝트 팀원을 결속시키명 장애를 극복하며 프로젝트를 수행해가는 "추진력"이 필요하다. 이러한 프로젝트를 성공적으로 수행하는 것이 데이터 분석 전문가이다.

빅 데이터 프로젝트의 기획 프로세스

빅 데이터 분석을 비즈니스 가치에 연결시키기 위해서는 분석 수행을 위한 기획이 중요하다. 비전 설정에서 활용 시나리오의 책정, 실용화 검증, 그리고 시스템 도입까지 빅 데이터의 전문가가 참여하여 기획 프로세스를 사전에 정립할 필요가 있다.

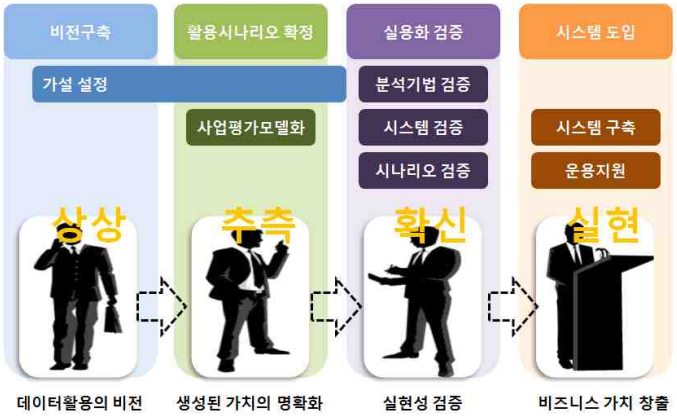



그림 20. 빅 데이터 프로젝트의 기획 프로세스

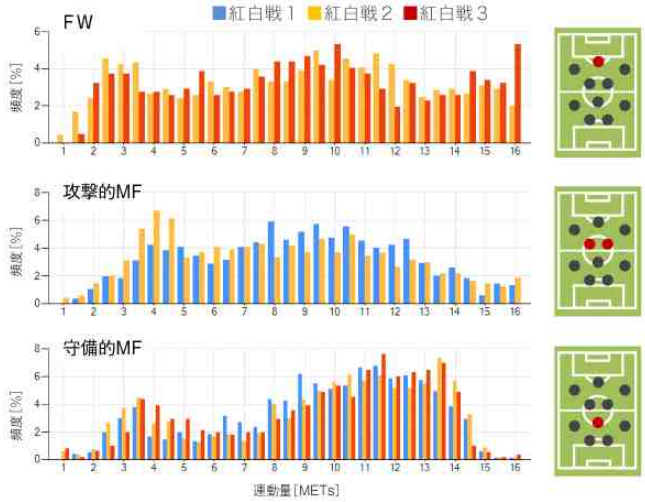
### *라이프 현미경을 활용한 로그분석 실증실험*

라이프 현미경을 사용한 경기 중 선수의 퍼포먼스를 분석

이번 프로젝트에서 사용한 센서는 히타치제작소의 라이프 현미경이다. 이것은 손목시계형 센서 네트워크형 단말기로 사람의 활동에 따른 3축 가속도를 24시간 365일 수집 가능하다. 이 단말기를 경기 중 선수에 장착하면 운동량, 보폭 수, 활동거리는 물론 대시와 조깅 등 주행 상태 분석을 통해 축구의 퍼포먼스 분석이 가능하게 된다. 포지션별 운동 패턴의 차이나, 운동량, 매 경기별 데이터를 비교 등을



상세하게 해석함으로써 각각의 특징을 명확하게 파악할 수 있다. 또한 프로파일링을 계속하면 1개월 전, 6개월 전, 1년 전 등 기간별 비교가 가능해 선수 개인의 변화와 성장을 시각적으로 확인할 수도 있다.




<빅 데이터 활용 축구 경기력 향상>

라이프 현미경(운동량 센서)

라이프 현미경을 사용하여 선수의 24시간 생활을 분석

또 하나의 실험은 선수에게 라이프 현미경을 생활 중에도 장착시켜 라이프 로그에서 평소 생활과 선수의 성과와의 관련성을 분석하는 것이다. 훈련이나 경기 이외의 생활은 아카데미 코치도 파악이 어렵다. 그러나 라이프 현미경을 선수에게 장착하며 두면 선수의 일상생활중의 운동량이나 운동의 강도, 수면상태까지 파악할 수 있다. 이렇게 분석 데이터가 축적되면 생활의 리듬에서 발생할 수 있는 잠재된 피로도 등이 사전에 파악되어, 부상을 방지하는 트레이닝 메뉴 개발과 조언 등에 활용 가능하게 된다.


<라이프 현미경을 사용하여 선수의 24시간 생활 분석 결과>

### *생활과 퍼포먼스와의 상관관계 발견*

코치도 몰랐던 선수의 컨디션 관리

이번 실증 실험에서 가시와 레이솔 아카데미에서는 수면과 퍼포먼스와 밀접히 관련되어 있는 것이 명확하게 파악되었다. 특히 코치를 비롯한 관계자를 놀라게 한 것은 선수가 훈련 전에 낮잠을 잤던 것은 코치의 조언이 아니라 훈련에서의 퍼포먼스를 최대로 하기 위해 선수가 자주적으로 실시하고 있는 것이었다. 고등학생이지만 가시와 레이솔 아카데미에 소속되어 프로선수로서의 강한 자각이 있음을 분석된 데이터로 증명한 것이다.

<선수의 컨디션 관리>

비즈니스로 연계되는 실증실험의 성과

라이프 로그 분석을 통하여 생활의 리듬에 잠재하는 경향을 파악하여 두면, 문제점이 발생하였을 경우에 정확하게 개선의 수단을 강구할 수 있다. 또한 비즈니스와 헬스케어, 간호나 교육 등 다양한 분야로의 전개가 가능하다.



## 콜센터의 영업실적에 영향을 주는 주요인 분석

### *콜센터 직원의 휴식시간 중 활성도와 영업실적과의 상관관계 분석*

실증실험의 개요

히타치는 (주)모시모시핫라인과 공동으로 콜 센터에서 전화를 사용한 영업활동에 "휴먼 빅 데이터 서비스"를 적용하였다. 콜 센터 업무 데이터와 행동 계측 시스템인 "비즈니스 현미경"을 사용해 측정된 사람의 행동 패턴 데이터에 분석 엔진 활용에 의한 복합 분석을 실시한 결과, 다수의 지표 속에서 휴식시간 중 직장의 활성도가 콜 센터의 수주율에 영향을 주고 있음을 구명하였다. 이러한 분석 결과를 활용하여 휴식 시간 스케줄을 개선함으로써 실적이 향상하는 것을 확인하였다.

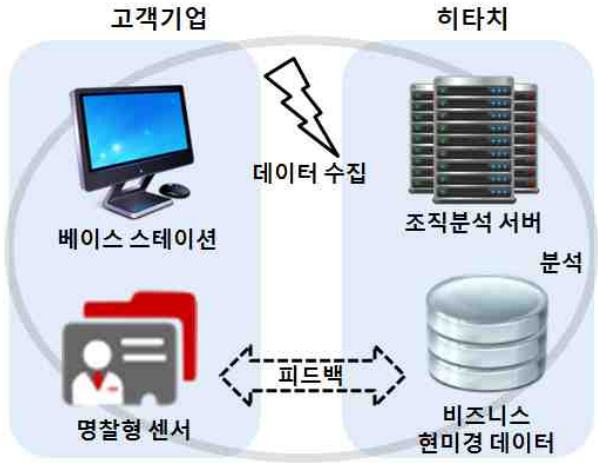



그림 21. 히타치에서의 빅 데이터 적용 체계

히타치는 지금까지 행동계측시스템인 "비즈니스 현미경"을 이용하여 인간의 행동을 100만일에 걸쳐 계측해 10조개의 빅 데이터를 축적함으로써 인간 행동 데이터의 해석 기술을 개발하였다.

이 시스템은 적외선 센서와 가속도 센서를 내장한 명찰형 센서를 이용해 누구누구와 언제, 몇분간 대면한 정보나 신체적인 활동을 데이터로서 취득, 해석하는 것이다. 이 시스템을 사용하여, 고객이나 파트너와 공동으로 영업 실적에 영향을 미치는 요인의 해명과 실적 향상 방안 수립에 활용하고 있다.

실증실험의 과제

콜센터에서는 스킬 향상을 목적으로 연수를 지속적으로 실시하고 있지만, 콜 센터 전체의 영업 성적이나 스킬 레벨이 제대로 향상되지 않고, 직장의 커뮤니케이션을 촉진하기 위한 다양한 방안을 실시하고 있지만 실적에 영향을 미치는 요인이 불확실하여 그 효과를 측정할 수 없는 콜센터의 과제로 제시되었다. 따라서 콜센터의 영업 실적에 미치는 요인이 무엇인가를 빅 데이터를 활용하여 규명하였다.

### *콜센터 영업실적의 요인 분석을 위한 빅 데이터 적용 방법*

사원 간 커뮤니케이션이나 직장의 활성도를 계측하여 영업 성적과의 관계를 정량적 분석

일본내의 2개소 콜센터(각 콜센터의 영업 담당자:51명, 79명)에서 각각 1개월간, 비즈니스 현미경을 이용하여 다음 항목을 측정했다.

- 영업 담당자간 및 상사와의 대면 정보
- 신체적 움직임

이들 데이터로부터 얻은 다수의 지표를 분석한 결과, "직장의 활성도"와 단위시간당 수주량인 "수주율"이 강한 상관관계에 있는 것이 규명되었다.

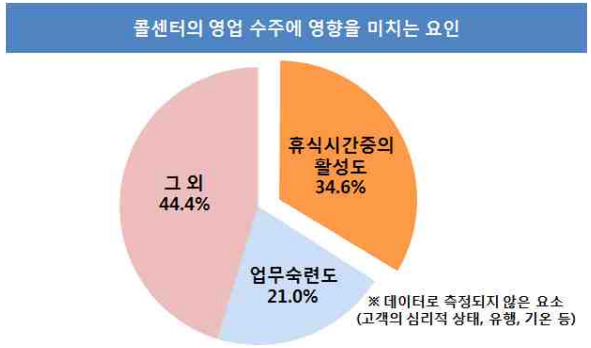



그림 22. 콜센터의 영업 수주 영향 요인 분석 결과

상기 그림은 콜센터간 수주율 격차와 일별 수주율 변동이 영업 스킬보다 휴식 시간 중 직장의 활성도[16]에 강한 영향을 받고 있음을 나타낸 것이다. 2개소의 콜센터의 수주율에 약 40%의 차이가 있었지만, 실적이 좋은 콜 센터는 다른 콜 센터에 비해 휴식 중 활성도도 약 40% 높다는 결과를 얻을 수 있었다.

16) 활성도 : 명찰형 센서에 내장한 가속도 센서로부터 얻는 신체적인 움직임을 나타내는 데이터에서 계산한다. 영업 담당자의 신체적 움직임 정도가 어떤 한계를 넘는 시간 비율로 계산하고, 그것을 직장 전체(전 영업 담당자)에 평균한 값을 직장의 활성도로 했다.



### *콜센터 영업실적의 요인 분석에 빅 데이터 적용에 의한 효과 방법*

콜센터의 1개 팀을 동 연령대의 담당자 4명으로 편성하여 휴식 시간이 일치하도록 작업 스케줄을 조정한 결과, 휴식 시간 중 직장의 활성도가 향상하였고, 수주율도 약 13% 향상되었다. 이것은 휴식 시간 중의 활성도 향상이 수주율 개선에 영향을 미침을 나타내는 것이다.

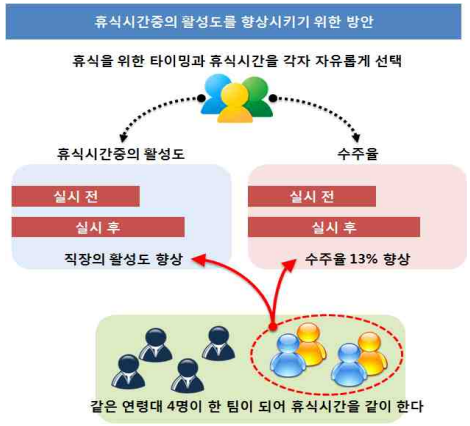



그림 23. 콜센터 영업실적 개선을 위한 방안

### *콜센터 영업실적의 요인 분석에 빅 데이터 분석에 적용한 기술*

휴먼 빅 데이터 서비스

기존의 경영 데이터와 센서가 자동 취득한 인간 행동 데이터를 통합 분석함으로써 다양한 현상, 인과 관계를 추출하여 경영 과제의 본질을 파악한다.

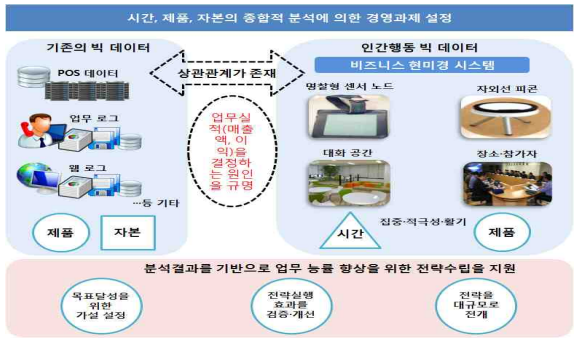



그림 24. 시간, 제품, 자본의 종합적 분석과 경영과제

비즈니스 현미경

명찰형 센서 노드와 적외선 신호를 이용하여 커뮤니케이션의 상황, 신체의 움직임 등을 계측하는 단말기이다. 누가 어디서 어떤 행동을 취한 것을 수치화하여 축적하기 때문에 조직 내의 커뮤니케이션 능력의 가시화가 가능하다.

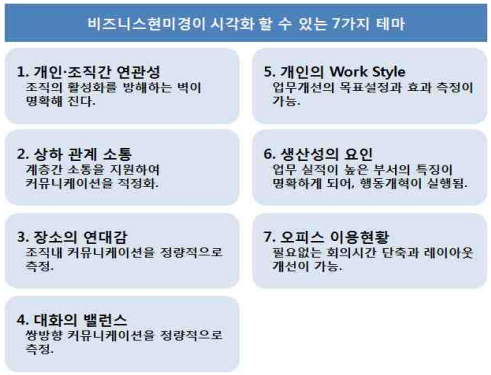




그림 25. 비즈니스 현미경 시각화

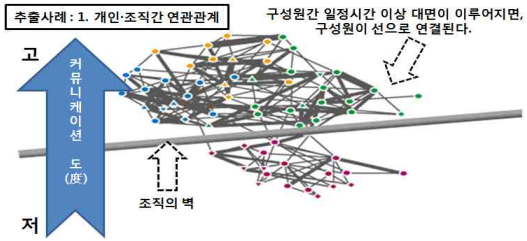



그림 26. 개인, 조직간 연관 관계



## 빅 데이터에 의한 가스터빈운전관리시스템

### *가스터빈센서에서 수집하는 데이터를 실시간으로 처리*

히타치는 세계의 화력발전소에 가스터빈을 납품하고 있으며, 이들 가스터빈에 설치한 센서로부터 가스터빈 상태 정보를 수집하고 있다. 가스터빈 1대당 약 200개의 센서를 설치하여 가스터빈운전관리시스템은 각각의 가동 상황을 일괄적으로 수집해서 분석·감시하고 있다.

과거의 장기간 축적된 운영 데이터를 활용하고 분석하여, 사고·고장의 전조를 사전에 검지하여 예방 보전에 활용하는 것은 가스터빈의 가동률 향상에 연결된다. 그러나 데이터양이 계속 증가함으로써 대량의 스토리지가 필요하게 되거나, 데이터 분석 처리에 소요되는 시간이 길어지면, 이렇게 축적된 데이터를 활용할 수 없게 되는 것이다. 따라서 가스터빈의 가동률을 향상시키기 위해서는 실시간으로 가동정보를 분석하는 것이 요구된다.

■ 각 사이트(고객)과 센터간을 접속하여 1일 1회, 1일분의 운전상황을 센서로 수집.
■ 과거의 각 사이트의 운전정보 및 다른 사이트의 운전정보를 다각적으로 분석하면서, 터빈의 운전상황을 파악

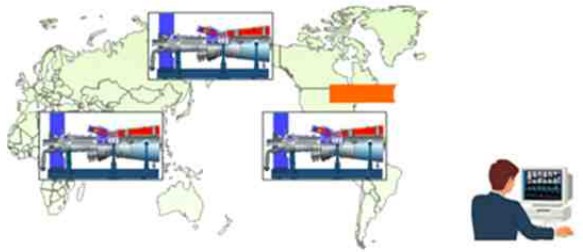


그림 27. 가스터빈 운전 관리 빅 데이터

### *가스터빈운전관리시스템에서의 과제*

가스터빈에 설치되는 센서의 양이 증가하고 데이터 수집의 주기가 점점 단기간화됨에 따라 운전정보의 데이터 량이 폭발적으로 증가하여 대량의 스토리지가 필요하게 되었다. 그리고 현재의 시스템에서는 운전정보를 RDBMS에 보관해 활용하고 있지만, 실시간 분석이 어렵고, 데이터 분석에 시간이 많이 소요되는 등 분석 업무의 작업 효율이 나쁘기 때문에, 다각적인 분석을 충분히 할 수 없는 문제점이 발생하고 있다.

### *데이터의 라이프사이클에 맞추어 운전정보의 감시, 축적, 분석*

이러한 기존의 가스터빈운전관리시스템의 문제점을 개선하기 위하여 운전정보의 라이프 사이클에 따른 데이터 관리기술을 적용하였다. 운전정보의 라이프사이클이란 운전정보를 감시, 축척, 분석의 프로세스 구성되며, 이를 기반으로 정확한 모니터링의 단계로 이어지는 사이클을 말한다.

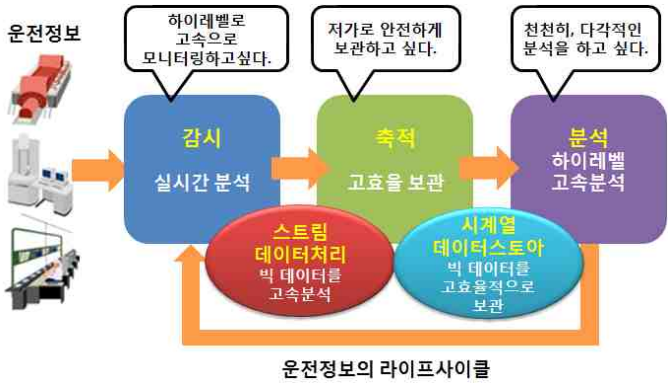



그림 28. 라이프 사이클 기반 운전 정보의 감시, 축적 분석

여기에는 스트리밍 데이터 처리기술과 시계열 데이터 스토어 기술을 적용하였다. 스트리밍 데이터 처리기술에는 빅 데이터 고속 분석기술이, 시계열 데이터 스토어기술은 빅 데이터의 고효율 저장을 가능하게 하였다. 스트리밍 데이터 처리 기술은 데이터를 축척한 후 처리하는 것이 아니라, 데이터가 발생할 때마다 메모리상에서 실시간으로 처리하는 기술로 빅 데이터를 고속으로 처리 가능하게 하는 것이다. 시계열 데이터 스토어 기술은 시계열 데이터를 특성에 맞춰 칼럼 단위로 압축해 격납함으로써 데이터 검색 및 분석 처리를 고속화하는 기술로 빅 데이터를 고효율으로 보관하는 기술이다.

### *빅 데이터를 활용한 가스터빈운전관리시스템의 도입 효과*

가스터빈운전관리시스템에 빅 데이터를 활용하여 운전관리를 하였을 경우에 빅 데이터 저장량의 삭감과 데이터 검색 시간의 단축이 가능하게 되었다.

스토리지량의 절감

- 운전정보(시계열 데이터)의 특성에 맞춰 칼럼 단위로 저장
- 운전정보를 1시간단위로 블로킹하여 압축 저장
- 스토리지량을 대폭적인 삭감이 가능



가동 정보 검색 시간을 단축

- 특징 인덱스에 의한 데이터 검색 속도 향상
- 운전정보를 블로킹해 격납함으로써, I/O 효율 향상
- 실제 운전 상황에 충분히 대응 가능한 데이터 검색 시간 구현

### *빅 데이터를 활용한 가스터빈운전관리시스템에 적용된 기술*

스트림 데이터 처리

현재 대부분의 기업에서 채용하고 있는 DBMS를 적용한 데이터 처리 방식은 분석 대상이 되는 정보를 일단 데이터베이스에 격납한 후, 배치(Batch) 처리로 일괄적으로 집계·분석하기 때문에 정보의 발생에서 집계·분석까지 시간 지연이 발생한다.

여기에 순차적으로 발생하는 데이터를 메모리상에서 실시간 분석하는 스트림 데이터 처리 기술이 적용된다. 스트림 데이터 처리 기술은 메모리에 사전에 등록된 시나리오에 따라 데이터의 집계·분석을 실시하는 것이다. 시나리오에는 처리하는 데이터의 조건과 처리 방법이 사전에 정의되어 있다. 이러한 데이터 집계·분석 처리 방식은 메모리상에서 실시되기 때문에 고속 데이터 처리가 실시간으로 실현 가능하게 된다.

- 인 메모리 처리, 차분계산처리에 의해 초고속 데이터 처리가 실현
- 시계열 데이터의 분석에 적합한 쿼리를 표준 제공
- 시나리오 베이스의 조건을 정의하는 것에 의해 어플리케이션의 수정이 불필요

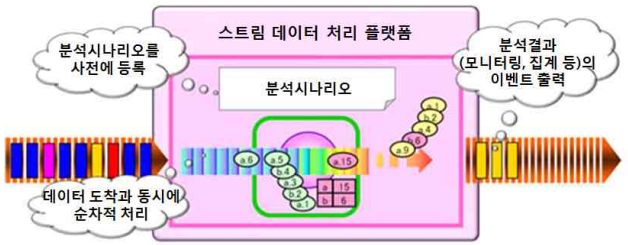



그림 29. 스트림 데이터 처리 플랫폼

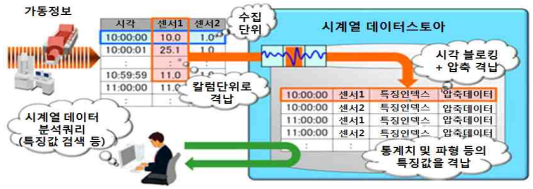



그림 30. 시계열 데이터 스토아

## 데이터 센터 공조감시시스템

### *히타치 그룹 내 데이터 센터 운용비용의 절감 문제*

최근 데이터 센터의 비즈니스 환경은 IT시스템의 대규모화·복잡화 속에서 서비스 레벨의 향상과 운용 코스트 저감이 중요 과제가 되고 있다. 히타치 그룹도 대규모의 데이터 센터를 운영하고 있어 데이터 센터의 비용 절감 문제가 그룹 내의 과제가 되고 있다.

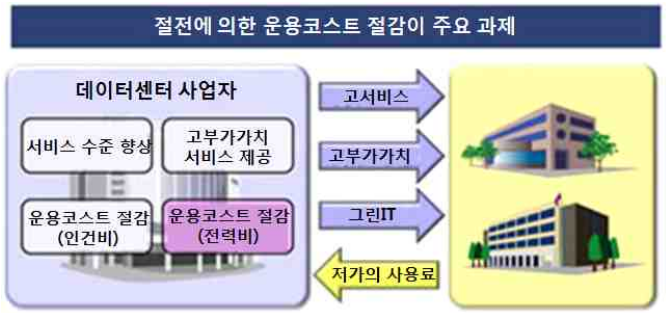



그림 31. 히타치의 데이터 센터 운용

### *데이터 센터 운용비용의 증가 원인*

서버 기기의 고밀도화에 따라 부분적으로 온도가 상승하고 있지만 정확하게 모니터링이 되지 않고 있으며, 서버실의 복잡한 레이아웃에 대응하기 위하여 "과냉각"로 대응하고 있기 때문에 운용 코스트(공조 비용) 증가로 데이터 센터내의 운영 비용이 증가하는 문제가 발생하고 있다.

### *데이터 센터 운용비용 증가의 해결 방안*

기본적으로 데이터 센터의 각각의 서버에 대한 개별 모니터링에 의한 공조 제어를 실시하여 실시간으로 이상을 검출하고 이상 상태에 대한 대응방안을 실시하는 것이다. 즉, 서버실 전체에 대한 모니터링뿐만 아니라, 랙 단위 및 짧은 주기별로 세밀한 공조 제어를 실시하여 과냉각을 방지하여, 공조비용의 절감을 구현하게 되었다.

국소적인 온도 상승 검지

저전력·방수형·무선기능이 내장된 온도 센서(히타치가 개발한 AirSenseU)"를 서버실의 여러 군데에 설치하여 국소적인 온도 상승을 사전에 검지한다.



고속 데이터 분석을 스트리밍 데이터 처리 기술로 실현

센서 데이터에서 순차적으로 발생하는 대량의 데이터를 스트림 데이터 처리하여 실시간으로 감시 및 경향 분석하여 이상 상태를 전조 단계에서 검지한다. 공조기에 나오는 공기의 배기 온도, 서버의 흡기 기온와의 상관 분석에 의해 센터내 열이 정체되는 현상이나 기류의 이상 상태를 검출한다.

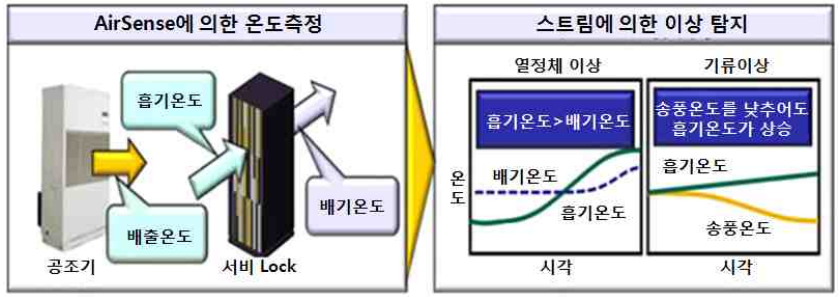



그림 32. 고속 데이터 분석 기반 스트림 이상 탐지

### *빅 데이터 적용 데이터 센터 공조관리시스템의 도입 효과*

빅 데이터 적용 데이터 센터 공조관리시스템을 도입하여 랙 단위의 흡기・배기 온도, 풍량의 실시간 감시를 통해 국소적인 온도 이상을 검지와 고효율의 공조 제어가 실현되어 데이터 센터의 공조비용의 절감이 가능하게 되었다.


그림 33. 랙 단위의 입도 작은 온도 감시와 공조 제어를 실현

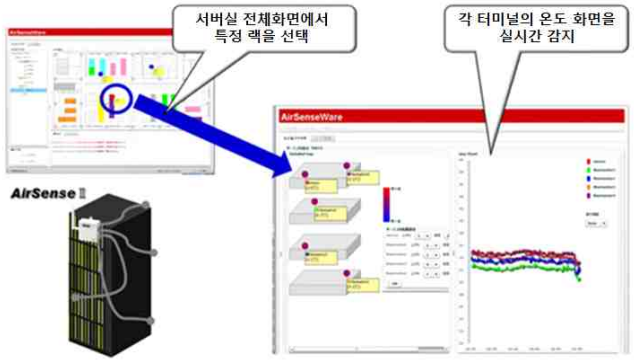



그림 34. 흡기・배기 온도 감시 시스템의 유저 화면

### *빅 데이터 적용 데이터 센터 공조관리시스템에 적용된 제품/기술*

히타치 AirSense

히타치 AirSense는 센서 네트워크 정보 시스템이다. 용도에 따른 다양한 센서(센서 노드)들이 무선으로 네트워크를 형성하여 비즈니스의 현장, 건물·시설이나 사람의 상황·환경(상태와 변화)을 적절하게 검지하여 발신할 수 있다.

그림 35. 히타치 AirSense

스트림 데이터 처리

스트림 데이터 처리 기술은  순차적으로 발생하는 데이터를 메모리상에서 실시간 분석하는 기술이다. 스트림 데이터 처리 기술은 메모리에 사전에 등록된 시나리오에 따라 데이터의 집계·분석을 실시하는 것이다. 시나리오에는 처리하는 데이터의 조건과 처리 방법이 사전에 정의되어 있다. 이러한 데이터 집계·분석 처리 방식은 메모리상에서 실시되기 때문에 고속 데이터 처리가 실시간으로 실현 가능하게 된다.

- 인 메모리 처리, 차분계산처리에 의해 초고속 데이터 처리가 실현
- 시계열 데이터의 분석에 적합한 쿼리를 표준 제공
- 시나리오 베이스의 조건을 정의하는 것에 의해 어플리케이션의 수정이 불필요



## 인텔의 빅 데이터 적용 사례

### *악성 코드 탐지*

사이버 범죄의 위협은 공격자와 공격 도구의 고도화와 함께 증가하고 있다. 보안 감시 기능의 목표는 사용자와 보안 직원이 대응 가능한 시간 내에 위협을 발견하는 것이다.

기존의 악성 코드 탐지의 가장 일반적인 방법은 바이러스 정의에 의한 파일 스캔 방식은  성 코드 수가 급증됨에 따라 유효성이 저하되고 있다. 더 좋은 방법은 항상 악성 코드의 움직임을 사전에 감지하여 악성 코드의 동작과 악성 코드의 출처를 분석하고 향후 악성 코드 가 발생 가능한 장소까지 예측하는 것이다. 이러한 고급 수준의 상세한 모니터링과 예측을 수행하려면 서버 동작을 항상 감시하고 시스템, 네트워크, 응용 프로그램의 각 레벨에서 이상이 없는지를 확인할 필요가 있다.

악성 코드의 위협을 나타내는 패턴은 종종 프록시, DNS, DHCP, VPN 등 각종 네트워크 로깅 및 서버 로그에 숨어 있지만 이러한 로그에는 대량의 데이터가 포함되어 있다. 이상 징후가 나타나는 방법은 일반적인 악성 코드 시그니처에서부터 악의적인 것으로 판명된 URL 과의 통신, 비정상적인 검색 등 비정상적인 활동과 동작의 다양한 글로벌 패턴에 이르기까지 다양한 형태가 존재한다. 따라서 이 분석에는 일련의 복잡한 절차가 필요하다. 비정상적인 활동을 감지하기 위해서는 여러 소스에서 얻은 데이터를 상호 연관시키고 정상적인 네트워크 활동과 패턴의 기준선을 정의하는 등의 단계를 수행해야 한다.

인텔은 규정된 시간 단위에 따라 이러한 이상을 특정시키기 위해 빅 데이터 기술을 사용하여 원시 비정형 데이터를 수집하고 그것을 구조화한 후 예측 분석 등의 통계 모델를 사용하여 비정상적인 활동 패턴을 감지한다.

이 개념은 악성 코드를 신속하게 식별하고 격리 가능하도록 비정상적인 동작을 실시간으로 파악하는 것을 목표로 하는 것이다. 수개월 또는 수년에 걸쳐 축적된 데이터를 수집하고 분석할 수 있으면, 보안 침해의 원인과 특성의 예측 정확도를 향상시켜, 보다 효과적인 예방 수단 및 예방 시스템의 도입을 가능하게 한다.

### *반도체 설계 검증*

반도체 설계에서는 설계를 실리콘에 실장하기 전에 광범위한 테스트를 수행할 필요가 있다. 이 테스트는 실리콘 실장의 각 단계에도 계속되며, 수백개의 센서가 초당 수천번의 샘플링 속도로 데이터를 수집한다. 이러한 광범위한 테스트의 결과로 많은 양의 데이터가 생성된다.

여기에 빅 데이터 분석 기술을 사용하면 테스트 프로세스의 최적화가 가능하게 된다. 수십억 컬럼의 구조화된 데이터와 구조화되지 않은 데이터를 분석하여 설계 프로세스의 신속화 및 양산까지의 기간 단축을 지원하여 빠른 시장 출시를 가능하

게 한다. 이 모델의 예로 커버레지라는 개념이 있다. 포스트 실리콘 검증의 세계에서는 칩을 출시하는 타이밍에 대한 명확한 규칙은 없다. 버그가 있는 칩을 출시하면 기업 평가에 큰 타격을 줄 수 있지만, 필요 이상으로 테스트를 실시하여 칩의 출시가 지연되어 매출 기회를 놓칠 수 있다.

커버레지의 개념은 이러한 극단적인 상황을 회피하는 것을 목적으로 하는 것이다. 프로세서가 테스트된 논리적 상태와 물리적 상태에 대한 데이터를 수집하면 테스트와 테스트 도구가 어떻게 기능하는가를 이해하여 칩을 시장에 투입 할 수 있는지를 판단할 수 있다.

또한 빅 데이터 분석을 통해 확인된 결함을 자동으로 클러스터링하여 재정렬 및 대량의 히스토리 테스트를 기반으로 근본 원인 분석을 수행하여 디버깅 프로세스를 지원할 수 있다. 수집된 대량의 데이터(샘플이 아닌)의 광범위한 분석을 통해 각 단계의 진행 상황을 포괄적으로 파악할 수 있으며, 설계 프로세스의 개선 및 합리화의 방법을 찾아 최종적으로 제품 자체의 개선을 할 수 있게 한다.

### *마켓 인텔리전스*

인텔처럼 세계적인 판매망과 글로벌 공급망을 가진 기업은 시시각각 변화하는 시장 환경을 예측하고 다음달, 6 개월후, 5~10 년 후에 어떤 일이 발생할 것인가를 정확하게 예측하는 것은 아주 중요하다. 글로벌 기업은 기상 트렌드, 세계 경제 데이터, 토론 포럼, 뉴스 사이트, 소셜 네트워크, Wiki, 트윗, 블로그 등 대량의 데이터 중에서 필요한 데이터를 선별해야 한다. 이러한 데이터를 바탕으로 정확하게 예측하고 판매 전략 수립, 경쟁사의 위험  평가, 소비자 행동 변화의 예측, 공급망 강화, 비즈니스 연속성 계획 수립이 가능하다. 즉, 다양하게 수집된 경영활동 데이터를 빅 데이터로 분석하여 경영활동에 적용하는 것으로 기업에서는 다음과 같은 결과를 달성을 목표로 한다.

- 세계 각 지역의 매출 예측을 개선하고 생산 수준을 미세하게 조정하고 보다 정확한 예측을 주주에게 제공한다.
- 발생 가능한 세계 규모의 사건에 기반한 시나리오 작성 및 테스트에 의하여 인텔의 시장, 공급망, 시장의 수요와 경쟁 업체의 도전에 대한 대응 능력에 미치는 영향을 평가한다.
- 인텔 제품의 신규 고객 및 새로운 기능을 발견한다.

### *추천 시스템*

콘텐츠의 양이 급격하게 증가하면 사용자는 자신의 관심에 적합한 정보를 효율적으로 찾기 위해서는 어떤 도움이 필요하다. 이러한 이유로 기업 내부 응용 프로그램과 외부 응용 프로그램 모두, 즉 기업 전사 차원에서 추천시스템 기반 서비스에 대한 수요가 증가하고 있다. Amazon과 Netflix가 고객에게 제공하는 시스템과



비슷한 추천시스템이 검색 및 탐색 시간을 단축하고 개인화된 타겟 결과를 제공함으로써 사용자를 지원한다. 따라서 생산성, 신뢰성, 전반적인 사용자 경험의 질을 향상시킨다.

확장성이 뛰어난 추천 시스템을 실현하려면 예측 분석 및 빅 데이터 분석에 대한 전문 지식이 필요하다. 이것은 많은 양의 기록 데이터에 대해 많은 리소스를 사용하는 복잡한 알고리즘을 실행해야 하기 때문이다.

이 시스템의 핵심은 빅 데이터 플랫폼 위에 2 층(오프라인 및 온라인) 아키텍처에서 재사용이 가능한 범용 추천 엔진 구축에 있다. 오프라인 컴포넌트는 추천 알고리즘의 핵심을 실행하는 배치 지향 프로세스이다. 이 구성 요소는 확장 가능한 환경에서 빅 데이터 처리가 가능하게 함으로서 모델을 확장 할 수 있도록 한 것이다. 온라인 구성 요소는 모든 서비스 요청에 대한 서비스 레이어 역할을 한다. 이 구성 요소는 오프라인 단계에서 계산에 관련 중간 계산을 로드하여 알고리즘의 마지막 단계를 실행하여 추천 사항을 생성한다. 또한 이 구성 요소는 컨텍스트 구성된 논리를 적용하여 요청된 컨텍스트에 따라 최종 추천 사항을 필터링하고 조정한다.

추천 서비스의 도입은 개인화된 콘텐츠를 고객에게 적시에 제공할 수 있도록 하는데 중요한 역할을 한다. 그리고 사내 직원들이 사내 응용 프로그램을 사용할 때도 생산성이 향상된다. 또한 경쟁력을 향상시키고 외부 고객이 자사 제품의 선택이 용이하게 되어 기업의 수익 향상에 기여한다.

## IBM의 빅 데이터 플랫폼 사례: IBM InfoSphere

IBM에서는 빅 데이터 환경에서의 정보 통합 및 거버넌스를 지원하는 플랫폼으로 InfoSphere을 제공하고 있으며, 이 플랫폼이 빅 데이터 처리를 위한 플랫폼의 한 형태로 볼 수 있다. 그림 4에 InfoShpere 구조도를 제시한다.

### *메타 데이터, 비즈니스 용어, 정책 관리*

데이터 통합 및 거버넌스를 실현하기 위한 필수 요소로 메타 데이터와 거버넌스 정책을 설정한다. 데이터 탐색, 메타 데이터 관리, 비즈니스 용어 정의 및 관리, 거버넌스 정책 정의 및 관리 등의 기능을 포함한다.

### *데이터 통합*

배치 데이터의 가공 및 마이그레이션, 실시간 리플리케이션(Data Replication), 데이터 통합 등과 같은 다양한 데이터 통합 기능을 의미한다.

### *데이터 품질*

전사적 관점에서 데이터 분석·표준화·검증·매칭을 수행하여 데이터 품질을 보장한다.

***마스터 데이터 관리***

다양한 데이터 도메인(고객, 제품, 계정, 위치, 참조 데이터 등)를 관리 기능을 의마하며, 모든 데이터 도메인 및 형식을 지원하고 필요에 따라 유연하게 사용자 정의 도메인을 정의 할 수 있도록 한다.

***데이터 라이프사이클 관리***

테스트 데이터의 생성에서부터 전사적 시스템 상의 데이터 삭제 및 보관 등 데이터 라이프 사이클을 관리하는 기능을 의미한다.

***프라이버시 보호 및 보안***

어플리케이션에 포함된 데이터 의 마스킹을 수행하여 중요한 데이터를 보호한다. 정보 저장소를 모니터링하여 데이터 위반을 방지하고 컴플라이언스를 확보한다.

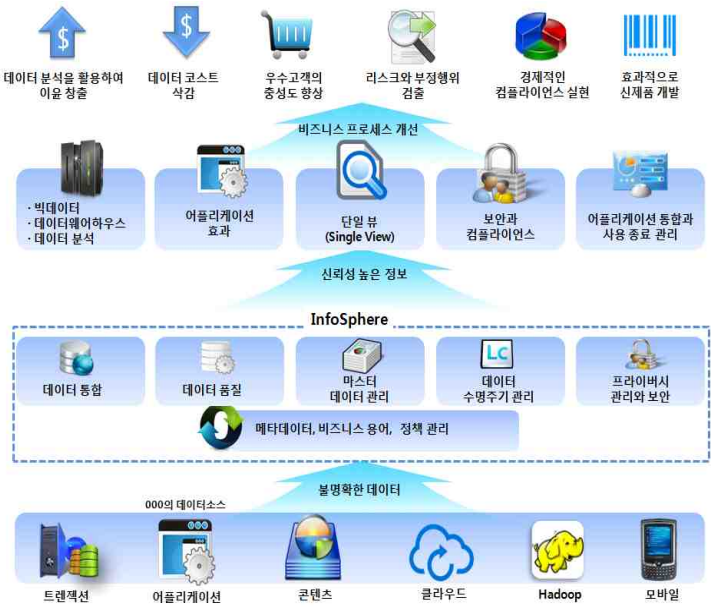



그림 36. 빅 데이터 플랫폼 구조도(IBM의 InfoSphere 사례)



# 빅 데이터 구현

## 빅 데이터 정보환경에서는 신뢰성이 높은 정보가 관건

최근 빅 데이터가 중심이 되는 정보환경으로 변화됨에 따라 정보의 통합 및 거버넌스의 중요성이 더욱 높아지고 있다. 데이터 분석 및 비즈니스 인텔리전스의 기본이 되는 것은 신뢰성이 높은 정보이며, 정보의 신뢰성이 낮으면 데이터에 의한 의사결정 지원이 될 수 없게 된다. 최근 데이터 활용의 중요성이 강조되는 정보환경은 다음과 같은 특징을 가지고 있어, 정보 통합과 거버넌스가 특히 중요하다.

### *다양한 데이터 소스에서 정보가 제공*

활용 대상이 되는 정보의 용량과 종류가 폭발적으로 증가하여 있어, 2020년에는 디지털 데이터의 용량이 현재의 44 배로 증가될 것으로 IDC가 예측하고 있다[17].

2 015년에는 트랜잭션 데이터, 소셜 데이터, 콘텐츠, 기계 데이터를 비롯한 약 80%의 데이터가"불명확한 데이터"가 될 것으로 예상되고 있다[18]. 즉, 데이터의 다양한 요소(데이터 소스, 품질, 데이터 원본의 신뢰도, 용도 등)가 불명확한 데이터가 급증하고 있는 것이다. 데이터 소스의 수가 폭발적으로 증가하여 데이터의 량이 급증하고 있지만, 신뢰할 수 있는 정보를 판별하는 것은 더욱 어렵게 되고 있다. 따라서 빅 데이터 시대의 정보 통합 및 거버넌스의 역할은 정보의 복잡성을 제거하고 복잡한 구조의 데이터 관리를 가능한 단일 데이터베이스 관리 수준으로 단순화하는 것에 목적이 있다.

### *유연성의 강화*

현재 기업의 데이터 관리측면에서 요구되고 있는 것은 정보 거버넌스 체제가 기업 활동에서 요구되는 높은 유연성에 즉시 대응 가능한 민첩성의 확보가 가능한가이다. 기업활동의 주체는 필요에 따라 새로운 어플리케이션, 데이터 형식, 보고서를 생성할 필요가 있다. 그 결과 기업의 비즈니스 룰을 벗어난 새로운 비즈니스 룰의 적용, 보다 복잡한 시스템의 구축, 거버넌스 관리를 벗어난 정보의 증대 등과 같은 현상이 발생하고 있다. 따라서, 이러한 복잡한 정보환경을 극복하기 위해서는 정보 통합 및 거버넌스 확보를 신속하게 실현하기 위해서는 정보 플랫폼을 활용하여 변화하는 사용자 요구에 신속하고 장기적으로 대응하며, 정보의 신뢰성을 유지하는 것이 요구된다.

### *확장성의 최대화*

모든 어플리케이션(특히, 정보 통합 및 거버넌스를 지원하는 어플리케이션)의 확장성을 극대화할 필요가 있다. 데이터 용량의 증대와 함께 처리해야 할 데이터 또한 급증하기 때문에 높은 성능을 제공하고, 정보 통합 및 거버넌스를 실현하는 기술의 확장성을 극대화하여 신뢰성높은 정보를 기업을 안정적으로 제공할 필요가 있다.

정보의 구축·공유·클렌징·통합·보호·유지 관리·폐기 방법을 내부적으로 통제함으로서 정보의 통합 및 거버넌스를 실현하여, 빅 데이터시대에서 불명확한 데이터를 신뢰성 높은 데이터로 변환할 수 있다. 여기에는 데이터 분석 어플리케이션과 비즈니스 어플리케이션은 불확실한 데이터의 증가와 데이터 용량 및 종류의 폭발적인 증가라는 두 가지 과제를 극복 할 수 있어, 빅 데이터 시대의 정보 통합 및 거버넌스 구현을 위하여 구현될 기술이 된다.

## 정보 품질의 중요성(에 내재된 핵심 가치)

본래 모든 정보(특히 정확한 정보)에는 다양한 가치가 포함되어 있다. 정보는 정보 분석을 통하여 정보의 가치를 더욱 극대화 할 수 있지만, 정보의 품질이 정보 분석으로 획득되는 통찰력의 가치를 결정한다. 품질이 낮은 데이터에서는 수준낮은 가치밖에 얻지 못하고, 양질의 데이터는 더 완전하고 정확한 분석 결과를 제공한다.

따라서, 품질이 낮은 데이터는 기업의 최종 의사결정과 성과에 심대한 악영향을 미친다. Data Warehousing Institute는 신뢰할 수 없는 "배드 데이터(Bad Data)"에 의해 미국의 기업의 경우 연간 6,000 억달러의 비용이 발생하고 있다고 추정하고 있다.[19] 향후 8년간 데이터의 용량이 50배 증가 할 것으로 예상하고 있어[20], 신뢰성이 낮고 막연한 배드 데이터가 발생시킬 비용도 동일한 정도로 증가 할 것으로 예상된다.

이러한 배드 데이터에 의한 문제에 대하여 적절한 조치를 취하지 않으면 다음과 같은 문제점이 발생할 수 있다. 신뢰할 수 없는 데이터로 발생하는 "비용"은 데이터 웨어하우스, 빅 데이터 분석 어플리케이션, 비즈니스 인텔리전스(BI), 그리고 전사적 비즈니스 어플리케이션 등 정보를 생성·활용하는 다른 모든 어플리케이션에도 영향을 미치고 있다. 신뢰할 수 없는 정보를 활용함으로서 투자 대비 효과를 얻을 수 없는 경우가 종종 발생한다. 또한 사용자가 시스템 에서 얻어진 정보와 통찰력을 신뢰할 수 없으며, 새로운 어플리케이션의 도입이 원활하게 진행되지 않을 수 있다.

정보의 양과 종류가 폭발적으로 급증하였지만 신뢰성 낮은 정보 또한 급증한 빅 데이터 정보환경에서 기업은 정보 거버넌스에 적극적으로 대응하지 않으면 상당한 비용 증대와 비즈니스 기회를 상실할 수 있다.

데이터에 내재된 가치를 극대화하려면 정보 거버넌스를 확보하는 것이 필요하다. 정보 거버넌스에 따라 빅 데이터 비용(중복된 데이터의 레코드 관리 비용, 비용을 발생시키는 부절절한 데이터 관리법, 기업의 명예를 훼손시키는 세큐러티 대응에

17) IDC. The Digital Universe Decade - Are You Ready? May 2010. http://www.ameinfo.com/231603.html
18) IBM Research. www.research.ibm.com/new-era-of-computing.shtml
19) Eckerson, Wayne W., Data Quality and the Bottom Line, Report of the Data Warehousing Institute, January 2002.
20) IDC. 2011 Digital Universe Study: Extracting Value from Chaos. June 2011.
http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

필요한 시간과 노력 등)을 절감할 수 있다

기업이 빅 데이터 분석를 통하여 목표를 달성하려면 고품질의 데이터가 필수 조건이다. 정보 통합 및 거버넌스를 실현한 기업만이 정보를 진정한 자산으로 관리할 수 있다.

## 빅 데이터는 기술이 아니라 현상이다.

빅 데이터와 함께 클라우드, 모바일, 소셜에 대응하는 기술이 소개되는 경우가 종종 있지만, 빅 데이터는 특정 기술을 가리키는 것이 아니라 현상을 가리키는 것으로 인식할 필요가 있다.

빅 데이터 정보환경에서는 데이터의 양이 급증(Volume)하고 있으며, 생성되는 데이터의 속도가 매우 빠르게(Speed) 진행되고 있으며, 생성된 데이터의 유형과 형식이 매우 다양화(Variety)되었지만, 데이터의 신뢰성의 불확실성 또한 급증하고 있는 것이다. 즉, 빅 데이터는 4 가지 관점 (용량, 속도, 유형, 신뢰성)에서 회사의 모든 어플리케이션에 영향을 미치고 있는 것이다(그림 1).


그림 41. 빅 데이터의 특성

## 빅 데이터의 기업 업무에 미치는 영향

빅 데이터 현상(대량의 다양한 종류의 데이터가 고속으로 제공되지만, 데이터 자체의 신뢰성 이 저하)은 기업의 모든 IT 시스템과 IT 프로젝트에 영향을 미친다. 빅 데이터가 Apache Hadoop 시스템과의 관계로만도 설명될 수 있지만, 대부분의 IT 담당자는 빅 데이터 현상이 시스템 모두에 영향을 미치고 새로운 수요가 발생되는 것을 인식하고 있다. 현재까지 관리 대상이 된 데이터와 다르게 다양한 특징을 가진 빅 데이터에 대하여 정보 통합 및 거버넌스를 적용하기 위한 관심은 도입되는 빅 데이터 관련 기술이 성숙됨에 따라 점점 높아지고 있다. 기술 검증 단계에서 실



제로 빅 데이터 운영을 개시하면 신뢰성 및 보안의 중요성이 한층 더 증가한다. 특히 빅 데이터를 관리하기 위해서는 종래와 같은 관계형 데이터 관리법이 아닌, 빅 데이터를 생성하는 데이터 소스를 관리하는 새로운 데이터 관리방안을 정립해야 한다.

빅 데이터 정보환경의 정보 통합 및 거버넌스를 구현하기 위한 과제로소 다음의 5 가지를 거론할 수 있으며, 모든 과제가 빅 데이터 현상의 영향을 받고 있다.

- 데이터웨어하우스 및 빅 데이터 분석에 신뢰성 높은 정보 제공
- 어플리케이션의 효율성 향상
- 전사적인 데이터 보호 및 안전성 확보하여 기업 컴플라이언스(Corporation Compliance) 확보
- 어플리케이션 통합 및 사용 종료 관리
- 신뢰성 높은 뷰(View)를 통한 데이터 분석에 의한 의사결정 지원

데이터웨어하우스 및 빅 데이터 분석에 신뢰성 높은 정보 제공

분석 어플리케이션이 정보를 처리·분석하려면 빅 데이터 플랫폼이 필요하다. 한편, 빅 데이터 엔진의 분석 엔진은 정보 통합 및 거버넌스를 실현하는 공통의 기반을 통해 안정적이고 신뢰성 높은 확실한 데이터를 제공함으로써 실제 기업 활동에 적용 가능한 정확한 통찰력을 제공하고, 이러한 통찰력을 다른 전사적 시스템에도 적용할 수 있게 된다.

어플리케이션의 효율성 향상

모든 어플리케이션(ERP, CRM, 트랜잭션 시스템, 데이터웨어하우스 등)에서 데이터가 급증가하고 있으며, 이러한 데이터의 증가는 성능과 효율성에 악영향을 미치며, 검색 처리 시간 의 증대, 프로세스의 지연, 정보시스템 사용자의 생산성 저하 등을 야기하고 있다. 또한 데이터 양이 증가하여 어플리케이션의 업그레이드가 더욱 복잡해져 결과적으로 시스템 작동 중지 시간의 증가 및 생산성 저하가 발생한다. 정보 통합 및 거버넌스를 정립하여 기업 컴플라이언스 규칙을 확립하고, 이에 따라 데이터를 관리함으로써 어플리케이션의 효율성을 향상시킬 수 있다. 또한, 테스트 데이터 관리를 합리화함으로써 어플리케이션 정상화에 소요되는 리드 타임을 단축할 수 있다.

전사적인 데이터 보호 및 안전성 확보하여 기업 컴플라이언스(Corporation Compliance) 확보

많은 시스템이 기밀 데이터를 공유하고, 보다 많은 사용자가 시스템에 액세스하고 있기 관계로 기밀 데이터의 마스킹 및 편집이 요구된다. Hadoop과 같은 최신 기술을 모니터링하여 데이터 위반을 제어할 필요가 있다. 또한 기업 컴플라이언스에 관한 리포팅 자동화를 통해 저비용 및 효율적인 기업 컴플라이언스을 확보할 필요가 있다. 정보 통합 및 거버넌스 확립을 통하여 데이터 보안을 유연하게 실현하

여 데이터를 언제 어디서나 사용되는 경우에도 영구적으로 데이터 보호가 가능하게 된다.

어플리케이션 통합 및 사용 종료 관리

어플리케이션을 효율화하려면 시스템 통합 및 사용 종료 관리를 통해 급증하는 시스템의 복잡성을 최소화할 필요가 있다. Fortune 1000대 기업의 대부분은 수천개 또는 경우에 따라서는 수만건의 어플리케이션과 정보 저장소를 보유하고 있다. 기업은 어플리케이션의 사용 종료를 위하여 필요한 조치(어플리케이션의 아카이브화와 사용 종료와 함께 데이터를 새로운 시스템으로 통합하는 것을 포함)를 수행해야 한다. 정보 통합 및 거버넌스가 확립되어 있으면 어플리케이션의 통합 및 사용 종료 관리를 신속하고 간단하게 기업 컴플라이언스을 확보할 필요가 있다.

신뢰성 높은 뷰(View)를 통한 데이터 분석에 의한 의사결정 지원

시스템이 증가함에 따라 정보가 분산된다. 기업은 수천의 정보 저장소 분산되어 있는 데이터의 통합 및 참조하여 분석을 실시하여 의사결정을 위한 통찰력을 추출해야 한다. 따라서 시스템의 복잡성이 증가하는 빅 데이터의 정보환경에서는 고객, 제품, 공급업체 등 중요한 비즈니스 주체별가 공통된 관점으로 사용 가능한 신뢰성 높은 단일 뷰를 구축할 필요가 있다.

이를 근거로 그림 2에 나타낸 바와 같이 빅 데이터 기술에는 다음의 사항이 포함될 필요가 있다.

- 데이터를 처리하고 분석하는 Hadoop 기반 시스템
- 스트리밍 데이터를 분석하는 스트림 컴퓨팅
- 대규모 병렬 처리를 위한 데이터웨어하우스 어플라이언스
- 빅 데이터 정보 저장소의 시각화 및 검색을 위한 정보 연계 기능에 기반한 검색 기능


그림 42. 빅 데이터 플랫폼(대량의 데이터 관리·분석·검색 기능이 포함)



## 빅 데이터와 거버넌스: 데이터 관리

데이터 거버넌스를 설정함에 있어 데이터 관리를 어떻게 할 것인가를 명확히 하는 것이 중요한 포인트이다. 데이터의 가치와 사용 목적은 빅 데이터 관리방안의 수립에 중요한 요소이다.

데이터에는 단기적인 가치밖에 가지지 않고, 효과가 즉시 상실되는 것도 있고, 장기적인 가치를 가지고 몇 년 동안 보존 할 필요가 있는 것도 있다. 그리고 데이터 사용 목적도 거버너스 설정에 있어 중요한 요인이 된다. 데이터에는 글로벌 데이터 및 익명 데이터의 분석용으로 사용되는 것도 있으면, 개별 레코드 수준의 분석용으로 사용되는 것도 있다. 정보 통합 및 거버넌스는 그림 36에서 나타낸 바와 같이 각각 4개의 영역(데이터 파악 영역, 데이터 보존 영역, 데이터 인식 영역, 데이터 장기적 보존 영역)에 따라 다르게 적용한다.

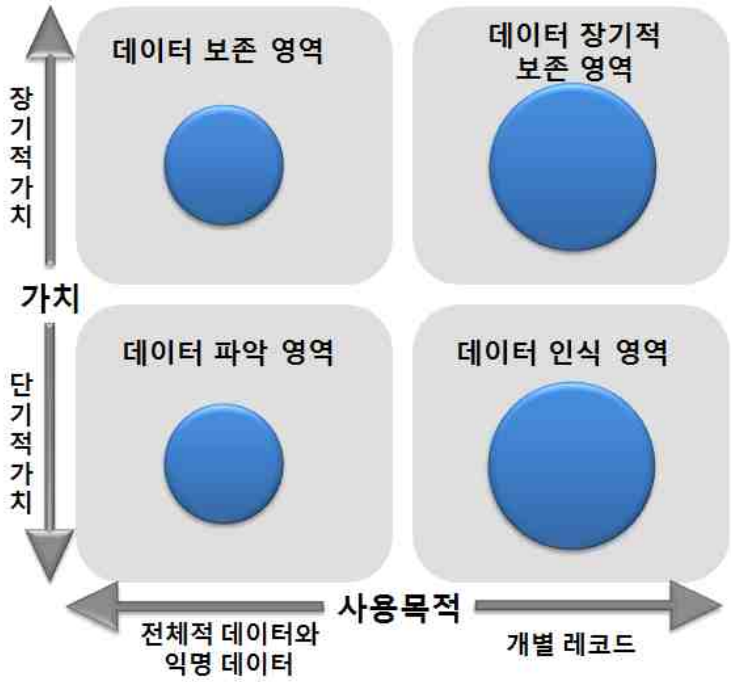



그림 43. 데이터 가치와 사용목적에 따른 4개의 데이터 영역

### *데이터 파악 영역*

이 영역에 해당되는 데이터 활용 방안은 데이터를 상호 결합하여 트렌드를 파악하는 것이다(소셜 미디어 분석을 통한 소비자 동향 파악 등). 이러한 데이터는 신속하게 축적해야 하며, 신선도가 생명이므로 신속하게 통합할 필요가 있다.

이 영역의 정보 통합 및 거버넌스는 데이터 제공, 데이터의 일관성 실현, 기밀 데이터 보호, 데이터의 시기적절한 삭제 또는 아카이브 관리이다. 데이터 라이프사이클 정책은 여러 조건 에 따라 적용되는 경우가 많다(예를 들면 지난달 데이터를 삭제하였지만 특정의 의사결정에 사용된 항목은 보관과 아카이브를 수행하는 등). 이러한 경우에는 데이터 저장 및 아카이브 정책이 데이터의 폭발적인 증가를 제어하기 위해 중요한 의미를 된다. 기밀 정보를 마스킹하여 개인 정보 보호 및 보안 확보와 동시에 의미 있는 정보를 제공할 필요가 있다. 데이터 의 일관성을 유지하기 위해 데이터 품질 정책을 적용할 수 있지만, 품질에 관한 모든 요구 사항을 엄격하게 충족할 필요는 없다.

### *데이터 저장 영역*

이 영역은 “데이터 파악 영역”과 유사하지만 이력 데이터에서 분석용으로 사용되는 데이터는 장기간에 걸쳐 저장된다. 일반적으로 데이터의 저장 기간이 길어지면 더 엄격한 거버넌스가 요구된다. 이러한 예로는 인구 동태 데이터의 분석과 재고 예측 등이 있다.

이 영역의 데이터 거버넌스는 데이터의 일관성을 향상시키는 것을 목적으로 하며, 데이터 의 품질에 대한 기능을 활용하여 데이터가 보다 일관된 형식으로 저장되도록 관리한다. 테스트 데이터의 관리 기능을 적용하여 시스템 및 업그레이드에 적절한 규모의 효율성과 보호 된 테스트 데이터를 제공 할 수 있다. 데이터 세트의 용량이 증가함으로써 기존의 배치 처리 중심의 데이터 추출·가공·로드 작업보다 많은 데이터 통합(다양한 리플리케이션(Replication)과 기능 연계를 포함)이 필요하다. 또한 데이터 증가를 억제하기 위하여 데이터 라이프사이클 관리가 중요한 기능이다.

정보 통합 및 거버넌스에 의해 데이터의 일관성을 향상시키며, 다양한 통합 기술을 통해 더 많은 데이터 소스의 데이터를 처리 할 수 있다.

### *데이터 인식 영역*

이 영역의 데이터는 주로 머신 데이터 분석 및 프로모션 분석 등에 사용되는 데이터로 데이터의 신선도가 높고 활용 시기가 매우 짧은 점에서 "데이터 파악 영역"과 유사한 특성을 가지지만, 광범위한 트렌드가 아닌 개별 레코드를 인식하는 것을 목적으로 하기 때문에 다른 특징을 가지고 있다.

이 영역에서는 데이터의 품질을 보장하는 범위가 넓어지기 때문에 거버넌스의 역할은 정확한 데이터의 제공이 데이터의 일관성 확보보다 중요한 의미를 가진다.



데이터 유효성 검사 및 매칭을 수행하여 마스터 데이터 관리(MDM; Master Data Management)을 통하여 분산된  데이터 세트에서 독자적인 마스터 엔티티를 구축한다. 데이터를 아카이브하여 데이터 증가를 억제하고 효과적으로 테스트 데이터를 관리하여 다양한 종류의 데이터 통합(배치 처리, 리플리케이션(Replication) 및 기능 연계)하는 것이 중요하다. 또한 데이터 분석의 리드 타임이 단기간이기 때문에 민첩성이 중요하다. 데이터 통합 및 거버넌스가 가속화할수록 기업은 보다 신속하게 데이터의 가치가 실현될 수 있다.

이 영역의 정보 통합 및 거버넌스를 실현하기 위해서는 민첩하게 데이터 통합·저장·아카이브하고, 기밀 데이터의 개별 레코드 속성의 마스킹을 위하여 보안 및 프라이버시 보호를 강화하고 개별 레코드의 인식을 위한 데이터 품질 및 마스터 데이터의 범위를 확대가 요구된다.

### *데이터의 장기적 저장 영역*

이 영역의 데이터는 기업의 중요한 의사결정을 포함하는 기업 어플리케이션에서 빅 데이터 분석 및 재무보고 시스템 등에 활용되며, 가장 엄격한 거버넌스 설정이 요구된다. 데이터에 대한 개별 레코드를 장기간 보존할 필요가 있기 때문이다.

개별 레코드를 저장하며, 레코드의 정확성을 최대한 극대화할 필요가 있다. 신뢰성 높은  데이터를 정확하게 유지하기 위해서는 MDM이 중요한 역할을 수행하며, 데이터 품질을 높임으로써 정보의 표준화와 검증을 실현한다. 데이터 라이프사이클 관리 대상이 되는 것은 전반적인 데이터(특정 데이터 블록을 제거하고 선택에 따라 보관)가 아닌 개별 레코드( 특정 고객 레코드를 보관)이다. 데이터 테이블뿐만 아니라 비즈니스 오브젝트에 대해서도 보관 및 추출을 수행해야 한다.

또한 이 영역에서는 보안과 프라이버시 보호가 큰 과제가 된다. 데이터 저장소를 모니터링하여 보안 위반이 발생하지 않는 것을 확인하며, 복수의 시스템이 기밀 데이터를 공유하기 위하여 중요한 데이터를 마스킹하거나 편집할 필요가 있다. 데이터 통합은 배치 처리, 리플리케이션(Replication) 및 연계 기능이 포함되며 개별 레코드가 정확하고 최신 상태로 유지되도록 관리해야 한다.

이상의 4개의 영역은 다양한 빅 데이터 활용에 있어 정보 통합 및 거버넌스의 개요를 나타낸 것이며, 이에 따라 빅 데이터 관련 기술을 적절히 작용할 필요가 있다.

## 빅 데이터 라이프사이클 관점에서의 주요 기술

### *빅 데이터의 이용 및 활용의 라이프사이클*

먼저 빅 데이터를 효율적으로 수집하며, 수집된 다양하고 방대한 데이터를 통합적 관점에서 저장·관리한다. 이러한 데이터에서 기업에 가치있는 데이터를 추출하고 분석·시뮬레이션를 수행하여 빅 데이터에 적용할 수 있도록 정보를 제공하며, 그 결

과를 실제 기업활동에  활용한다. 그림 37는 빅 데이터 이용 및 활용의 라이프사이클을 나타낸 것이다.

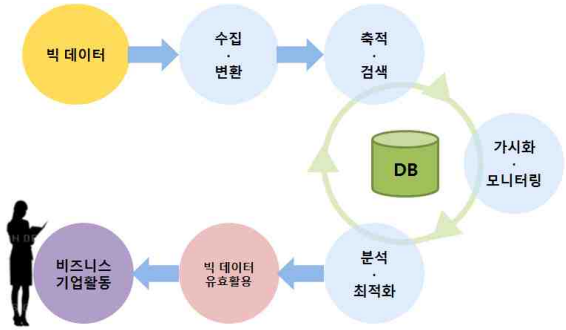



그림 44. 빅 데이터 이용 및 활용의 라이프사이클

***기업내 데이터의 활용상의 과제 및 해결방안***

데이터가 빅 데이터시대가 되어 기업에서 활용할 수 있는 데이터가 급증하였지만, 실질적으로 기업활동에 데이터를 활용하고자 하였을 경우에는 그림 6에 나타낸 것과 같은 과제가 있으며, 과제에 대한 해결방안으로 데이터 가시화, 데이터 가상화, 데이터 병렬화, 데이터 추상화 방안이 있다.

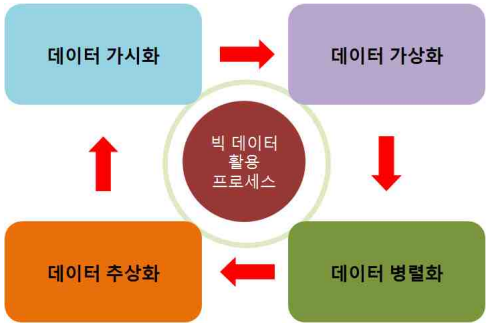



그림 45. 기업내 데이터의 활용상의 과제 및 해결방안

그리고 그림 38에 제시된 4가지 빅 데이터 해결방안의 프로세스를 그림 7에 나타낸다.



**데이터 가시화**: 현장 데이터를 수집하여 기존 데이터와 연계하여 업무 및 데이터와 관련성을 시각화하는 기술

**데이터 가상화**: 데이터의 저장 위치, 구조, 관련성, 내용 등을 통합된 데이터 관리를 실현하는 기술

**데이터 병렬화**: 데이터의 병렬 분할과 실행을 저장 장치의 병렬성에 자동으로 조정하고 실행하는 기술

**데이터 추상화**: 데이터 개요, 상호 관계, 잠재 구조 등을 분석 및 추출하여 데이터를 정보로 승화하는 기술

그리고 각 빅 데이터 기술별로 각 기술별 구현 기술을 정리하면 그림 8과 같다.

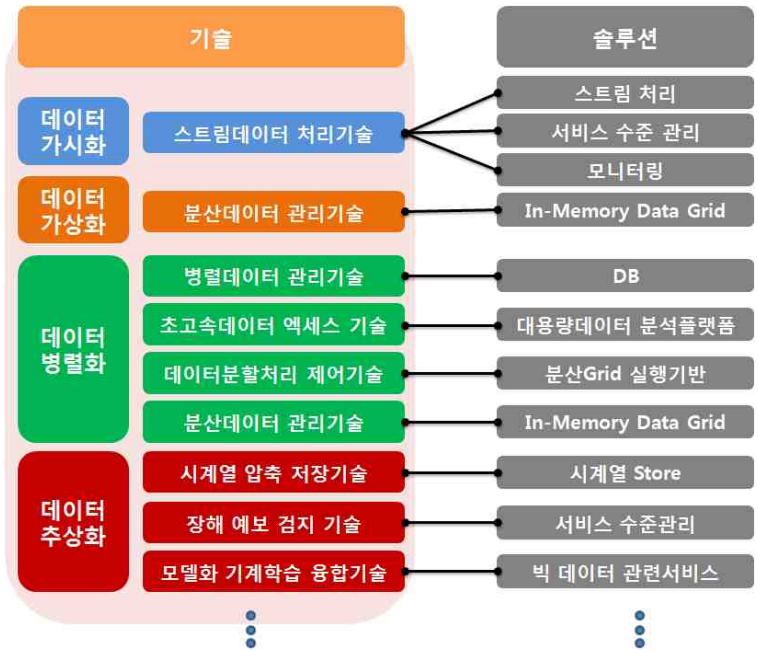



그림 46. 빅 데이터 기술별 구현 기술(예)

# 빅 데이터 활용 분야

## 빅 데이터 활용 분야

빅 데이터는 다양한 분야에서 불가능한 가능으로 전환시킬 수 있는 비전을 제시하고 있으며, 정보 기술과 빅 데이터의 결합은 조직 내 연구 개발 성과 및 속도를 향상시킨다. 빅 데이터는 다양한 분야에서 활용할 수 있는데, 맥킨지가 제시한 빅 데이터를 활용하여 가장 큰 효과를 얻을 수 있는 분야는 아래 그림과 같다.

표 10. 맥킨지에서 제시한 빅 데이터 활용 분야

| 도메인 | 분석 대상 데이터 | 예상 효과 |
|---|---|---|
| 미국의 의료 산업 | 제약사 연구 개발 데이터, 환자 치료·임상 데이터, 의료 산업의 비용 데이터 | 연간 $3조로 0.7% 생산성 향상 |
| 유럽의 공공 행정 | 정부의 행정 업무에서 발생하는 데이터 | 연간 $4.1조로, 0.5% 생산성 향상 |
| 소매업 | 고객의 거래 데이터, 구매 경향 | $1조+서비스 업자 수익<br>$7조 소비자 이익 |
| 제조업 | 고객 취향 데이터, 수요 예측 데이터,<br>제조 과정 데이터, 센서 활용 데이터 | 60% 마진 증가<br>0.5~1.0% 생산성 증가 |
| 개인 위치 데이터 | 개인과 차량의 위치 데이터 | 개발 및 조립 비용 50% 감소<br>운전 자본 7% 감소 |

빅 데이터의 활용 분야를 공공, 과학, 의료, 도소매, 제조, 정보 통신 등 여섯 개로 구분하여 그 특징을 정리하면 다음과 같다.

### *공공분야*

공공 분야는 국가적 차원에서 방대한 양의 데이터를 바탕으로 수자원 관리, 스마트 그리드, 재난 방재 영역 등을 포괄적으로 포함한다. 기업적인 이익은 낮으나 국가적 효용 가치는 매우 높은 영역이며, 이 분야에 빅 데이터를 활용하면 투명성과 개방성, 높은 수준의 분석으로 생산성 향상과 운영 효율화, 공공 분야의 경쟁력 강화로 국가 시스템 전체의 경쟁력을 향상시킬 수 있다.

대표적인 예는 싱가포르 정부가 빈번히 발생하는 테러 및 전염병 등 불확실한 미래에 대비하려고 2004년부터 추진하고 있는 구가 위험 관리 시스템이다.

### *과학 분야*

대용량 지리, 기상, 우주과학 분야 등에서 각 기관별로 과학 데이터를 처리하고 있으나 체계적인 수집 및 배포 경로가 미흡하며, 또한 데이터 생산량에 비해 데이터 활용, 재사용, 보존도 원활하게 진행되지 않음. 산발적으로 흩어진 과학 데이터를 국가 차원에서 수집, 가공, 유통, 재활용할 수 있는 기반을 마련하여 활용하면 환경, 기후, 해양, 등에서 일어나는 문제를 효율적으로 해결할 수 있다.

### *의료 분야*

의료 기록의 전자화, 병원 간 연구 데이터 공유로 빅 데이터 도입과 활용이 확대되고 있는데, 이는 공공 분야와도 밀접한 연관성이 있으며, 특히 미국에서는 빅 데이터를 사용하여 의료 분야에서 직간접적으로 개선한 비용이 연간 약 3300억 달러에 이를 것으로 전망하고 있다. 미국국립보건원(NIH)은 의약품 검색 서비스인 '필박스'로 주요 관리 질병의 분포, 연도별 증가 등 통계치를 확보해 연간 5000만 달러의 비용 절감 효과를 보고 있다. 이 밖에도 IBM과 미국 건강 보험 회사 웰포인트는 인공지능 컴퓨터 시스템 왓슨(Watson)의 빅 데이터 분석 능력을 활용하여 의사와 다른 의료진이 진단과 환자 치료에 이용할 수 있는 의료 데이터 활용도를 향상시키는 서비스를 제공 중이다.

### *도소매 분야*

도소매 분야에서는 이미 비즈니스 인텔리전스(BI)와 고객 관계 관리(CRM) 등으로 데이터를 분석하고 활용해 왔기 때문에 빅 데이터를 연속성 있게 활용할 수 있는 것이 중요하며, 빅 데이터 분석으로 시장의 정적, 동적 요인을 더욱 정밀하게 분석하여 이에 따른 적시 수요 예측 및 선제적 경영 지원에 초점을 두고 있다.

### 제조 분야

제조는 보유 데이터양이 많고, 불량품 개선비용등 적용 효과를 계량화하여 빅 데이터의 유용성을 손쉽게 확인할 수 있는 분야이다. 기존에도 품질 관리 및 보증이 진행되고는 있으나, 인적 교육에 기반을 둔 현재의 방법론보다는 데이터 기반 방법론을 사용하면 안정적으로 품질 관리 및 보증이 가능하다. 자동차 업계는 빈번히 발생하는 리콜 사태를 사전에 예방하고, 자동차 결함을 제품 출시 후 단기간 내에 파악하려고 빅 데이터 분석을 도입하고 있다.

### 정보 통신 분야

정보 통신 분야는 이동 통신의 발전과 개인 단말기의 증가로 생성된 디지털 공간의 개인 데이터를 기반으로 목표 마케팅, 개인화 서비스를 확대하며, 다른 분야보다도 인터페이스 측면이 강조된다. 현재 개인의 위치 정보 수준에서 진행되고 있는 서비스들은 점차 사용자의 행동 패턴, 이력, 선호도, 주변 상황에 따라 적절한 지식을 제공하는 서비스 형태로 진화할 것이다. 애플의 시리(Siri)는 사용자의 스마트폰을 음성으로 작동시키고 대화할 수 있도록 빅 데이터 분석 과 편리한 사용자 인터페이스를 결합한 대표적인 서비스이다.

빅 데이터 기술이 발전하면 최근의 다변화된 현대 사회를 더욱 정확하게 예측할 수 있고, 효율적으로 작동하도록 정보를 제공할 수 있음. 또한 개인화된 구성원들에게 맞춤형 정보를 제공, 관리, 분석할 수 있게 하며, 과거에는 불가능했던 기술까지도 발전시킬 수 있다. 대표적인 예는 다음의 이상 현상 감지, 고객 이탈을 사전에 감지한 T-Mobile, 위키리크스 데이터 분석을 이용한 효과적인 정보 제공, 아마존닷컴의 추천 상품 표시와 구글 및 페이스북의 맞춤형 광고 등을 들 수 있다.



### *이상 현상 감지*

이상 현상 감지는 업무에서 발생하는 다양한 이벤트를 기록하여 정상 상태, 비정상 상태를 표시하는 패턴을 파악하고, 이 패턴을 기초로 새로운 이벤트가 발생했을 때 이상 현상 여부를 판단하는 것이다. 예를 들어, VISA 신용카드사는 시스템 로그의 패턴을 분석하여 내부 범죄 등 부정행위를 알아내어 카드 부정사용을 방지한다. 그리고 HP는 시큐리티, 이벤트 관리 솔루션인 ArcSight를 도입하여 출입 관리 시스템, 네트워크, 업무용 응용 프로그램, 데이터베이스 로그 등을 종합적으로 분석하여 내부 통제를 위반한 내부 부정행위를 적발한 적이 있다.

이상 현상 감지를 부정, 범죄 분석뿐만 아니라 마케팅 부야에도 활용할 수 있음. 제품과 서비스에 변심한 고객을 감지하여 고객 이탈 방지를 막는 데 활용하는 것이다. 콜센터의 고객 목소리를 음성 인식 소프트웨어로 텍스트화하여 분석할 수 있으며, 스팸 메일 필터링에 걸리는 특정 벡터를 하둡으로 추출하여 고객 마케팅에 활용할 수도 있다. 또한 의료, 간호, 분야에서도 의료인들이 감지하기 어려운 이상 현상을 신속하게 감지하는 데 활용한다. 캐나다 온타리오 공과대학은 집중치료실(ICU)에 있는 약 100명의 화우 아동을 대상으로 심전도, 심박수, 혈압 등 16종류의 검사 결과 수치를 수집, 분석하여 패턴을 도출해 신생아 이상 징후의 감지에 활용하고 있다.

### *고객 이탈을 사전에 감지한 T-Mobile*

미국의 T-Mobile은 자사가 보유한 빅 데이터를 분석해 고객의 통신사 전환 위험을 감지하는 시스템을 운영하고 있으며, T-Mobile에서는 3000만 명이 넘는 가입자에게서 매일 179억 건 이상의 통화 및 송수신 내역을 담은 빅 데이터를 발생하고 있다. 빅 데이터 분석으로 다른 통시사로 회선을 옮긴 고객이 사전에 보인 이용 패턴을 감지하고, 이를 실시간으로 포착해 내는 시스템을 구축했다. 또한 소셜 네트워크를 분석하여 이탈 징후를 보이는 영향력이 큰 고객을 따라서 지인들이 동반 이탈하는 현상을 발견하고, 이탈 징후를 보이는 고객에게 맞춤형 추가 혜택을 제공하여 이탈을 방지함. 이 시스템으로 고객 이탈이 절반 수준으로 떨어졌다.

### *위키리크스 데이터 분석으로 효과적인 전술 정보 제공*

드루 콘웨어(NYU 박사 과정)는 위키리크스에 저장된 테라바이트급의 핵심 데이터를 분석해 미국과 아프가니스탄 연합군의 병력 활동 동향을 알아냈으며, 데이터 분석에 R 언어를 사용했고, 아프가니스탄 주요 다섯 곳을 적, 중립, 동맹 지역으로 나누어 정보를 분류한 후 각 지역에서 어떤 활동이 주로 일어나는지 패턴을 분석했다. 그리고 이 결과로 탈레반의 활동이 어느 지역에서 많이 일어나는지, 미국과 동

맹을 맺은 지역이 어딘지 쉽게 파악할 수 있었음. 또한 시간 흐름에 따라 아프가니스탄에서 전쟁 양상이 어떻게 진행되는 지도 확인할 수 있었다.

***아마존닷컴의 추천 상품 표시, 구글 및 페이스북의 맞춤형 광고***

아마존닷컴은 모든 고객의 구매 내역을 데이터베이스에 기록하고, 이 기록을 분석해 소비자의 소비 취향과 관심사를 파악하였으며, 이를 바탕으로 고객별로 추천 상품을 표시함. 고객별로 취미나 독서 성향에 맞는 상품을 메일이나 웹 사이트에서 자동으로 제시하는 것이다. 구글 및 페이스북도 사용자의 검색 조건, 나아가 사진가 동영상 같은 비정형 데이터 사용을 즉각 처리하여 사용자에게 맞춤형 광고를 제공하는 등 빅 데이터의 활용을 증대시키고 있다.

# 빅 데이터 시대를 위한 해결 과제

## 빅 데이터에 활용에 관련된 해결해야 한 과제

### *데이터의 생성·수집·축적에서의 과제*

o 동종의 데이터가 생성되는 센서의 관리자가 다양한 경우에 데이터 수집 방법 정립

o 다양·다양한 센서에서의 데이터 수집에 있어 센서 설계 및 통계적 기법과의 연계 및 ICT 인력과 통계 분석 전문가의 활용 방안 정립

o 데이터 수집에 있어 GPS의 보급과 데이터 입력에 대한 사용자 인터페이스 설계 등과 같은 이용자로부터의 데이터 수집 방안 정립

o 개인정보를 이용하는 서비스 제공에 있어서 계약 약관에 의한 동의 취득 및 중요 사항에 대한 규정 정립

o 센서 등으로 생성된 데이터를 실시간으로 수집·처리하기 위한 대역폭 및 지연 시간 등에 대한 네트워크 및 처리 시스템의 인프라 환경 정비

o 센서를 네트워크화하여 데이터 수집 대상의 규모에 따른 비용 효과와 센서를 설치에 따른 무선 충돌 방지 방안 설정

o 기업 등 이용자의 데이터 활용의 중요성에 대한 인식 및 이용자 ID와 연계한 데이터 정비·관리 방안 정립

o 여러 주체 에서 다종 다량의 데이터를 클라우드에서 사용될 경우에 적절한 데이터 관리 방안 정립

### *데이터의 유통·연계에서의 과제*

o 정보를 정확하게 상시 보내는 경우의 방송 매체 등의 전국적인 인프라 정비

o 기하급수적으로 증가하는 데이터 트래픽 등 빅 데이터를 지원하는 기반이 되는 서버, 스토리지 및 네트워크 등에 대해 소프트웨어로 프로세스를 제어하는 기술 및 Hadoop 등의 각종 스토리지 기술 등의 활용과 소비전력 등을 고려한 아키텍처 등의 구성 방안 정립

o 휴대폰 네트워크를 활용한 데이터의 송수신이나 서비스 판매 플랫폼으로 휴대폰 활용에 있어 스마트폰에 활용방안 정립

o 위치정보 등의 정보 자체와 그것을 기반으로 한 특정 사실만의 제공 등, 정보의 종류에 따른 위험과 기회를 고려하여 정보사업자간의 역할 및 정보관리 방안 정립

o 지진 재해시, 자동으로 다양한 사람이 서버에 전송한 데이터를 서버간에 연계·가공 하여 필요한 정보를 관계자에게 전달하는 공적 플랫폼과 체제 정립

o 민간 기업에서 수집된 데이터를 평상시에 공개하여 유통·연계하는 경우에 있어 데이터 형식의 표준화 및 메타 데이터 정의, 정보 제공자의 이익을 배려한 비즈니스 모델 설정



o 각 분야의 규칙과 정보 활용 능력 등 의 차이 등을 고려한 정보 공유를 위한 입력 및 정보 표시 설정

### *데이터의 활용상의 과제*

o 해외 기업의 국내에서의 서비스 제공 상황을 감안한 이용자와의 명시적 계약 체결 및 통계적 익명화 처리를 통한 재해 대응 등과 같은 공공 목적의 통신 로그 사용 방안 정립

o 통신 서비스의 제공에 있어 위치 등록 정보 등의 운용 데이터 비식별화처리기술과 공공분야에서의 활용 방안 정립

o 의료 분야 등에 있어서 민감한 정보 취급법 등 개인정보 관리 방안 정립

o EU의 데이터 보호 지침 개정과 미국의 Do Not Track 등 세계적인 국제 동향에 대한 대응 방안 정립

o 개인 식별성이 없는 정보가 계속적으로 유통되는 가운데 개인 식별을 가질 가능성이나 사생활을 침해 할 수 있는 라이프 로그 활용 서비스의 제공에 있어 투명성 확보 및 이용자 참여 기회 제공 등의 개인 정보 보호에 관란 방안 정립

o 데이터 내용의 암호화 및 ID 처리 방법과 k-익명화에 관한 기술 연구 등의 개인 정보 데이터에 대한 은닉화 기술 개발 및 운영 방안 설정

o 잘못된 정보가 이용자에게 전송되는 경우에 있어 정보의 수신 단말기의 갱신과 계약상의 책임 감면 조항 규정 등 이용자의 판단과 제공측의 책임 등 책임 소재에 대한 규정 설정

o 데이터의 정확성이나 관측기의 정확성등이 제도상 규정되어 있는 기상 데이터 등을 인터넷에 공개함으로써 다른 분야와 다른 목적으로 활용하는 경우에 대비한 데이터 정확성 확보 방안 설정

o 실생활 서비스 제공에서 발생되는 로그 데이터 수집 및 다양한 데이터를 조합하여 활용 가능한 시스템 구축 방안 정립

o 재해에 관련된 개인의 다양한 정보에 있어 개인 정보 및 프라이버시, 저작권이 보호되면서 정보가 유용하게 사용될 수 있도록 제공 형식이나 인터페이스 설정 등 정부 및 지방자치단체가 보유한 데이터 공개 방안 설정

o 공공 데이터의 데이터 소재의 명확화 및 데이터 수집시의 형식 등의 프로토콜 설정

o Hadoop 등과 같은 오픈소스 소프트웨어에 관한 운영 커뮤니티가 매수된 경우에 운영 주체의 부재에 따른 대응 방안 정립

o 호스트 컴퓨터에 올리기 전에 메타 데이터화에 대한 포맷 변환 등, 향후 실시간 데이터 처리를 위한 고성능의 새로운 데이터 처리 인프라 및 분석 기술 확립

o 현재 이용 목적별로 다양한 데이터 형식에 의한 응용 프로그램 중심의 데이터 활용과 달리 향후 M2M 등의 다양한 데이터가 활용되는 경우에 개별 응용 프로그램을 기반으로 한 보안 및 프라이버시 보호 등의 응용 프로그램의 수평부분의 표준화 문제 정립

o 수집된 데이터의 귀속 대상 및 해당 데이터에 대한 제3자에 의한 무단 사용에 대한 대응방안 정립

o 영업 비밀로서의 관리방안 및 기밀유지계약에 의한 개별기업에 제공이나 제3자에 대한 제공 등, 저작물로 보호되지 않는 데이터 분석 결과에 관리 방안 정립

o 공학·경제학·의학 등 다양한 분야의 비정형 데이터의 활용에 관한 수학적·통계학적· 법학지식과 경영학 지식을 구비한 인재 확보· 육성

o 행정 기관이 민간에서 구입하여 활용하는 데이터의 관리 방안

o 데이터의 오남용··부적적한 이용을 발견하기 위한 알고리즘의 개발

o 다량 및 다양한 특성을 가진 빅 데이터에서 가치있는 정보를 찾기 위한 분석 도구, 데이터 분석을 위한 노하우의 지속적인 축적, 그리고 적절한 데이터 검색을 위한 현장 비즈니스 정보 수집 방법과 데이터 분석을 위한 가설 구축· 검증· 데이터 검색 방법 개발

o 빅 데이터 활용의 성공사례 및 해외 성공 사례의 공유를 통한 데이터 활용의 중요성 인식 증대

o 기업이 보유한 방대한 데이터를 신규 사업에 활용을 위한 타 업종과의 제휴 모델 정립. 예를 들면 사회 인프라 및 의료 분야 등 신산업 창출 및 사업 영역 확대가 예상되는 분야에 투자를 촉진해서 새로운 시장을 창출 기여할 수 있도록 데이터 융합을 촉진하기 위한 시스템 정비

o 디지털기술과 물리적 제품과의 융합을 중심으로 Google이나 Facebook 등이 현재 주력으로 하지 않는 분야 발굴로 글로벌 경쟁력 확보

o 개인 정보의 활용에 따른 이익과 손실의 밸런스 문제 정립

o 실생활에서 수집한 데이터를 이용한 분석 결과를 비즈니스 또는 실생활에 적용을 지원하기 Open Innovation 플랫폼 구축

o 해외 성공적인 프로젝트에 적용되어 검증된 기술을 국내 적용 및 공적 효과의 검증을 위한 빅 데이터 실증사업 실시

o 스마트폰 보급과 테러 대책으로서의 소셜 모니터링을 위한 웹 페이지의 다언어화에 대한 통계적 자연언어처리를 통한 다국어 분석방안 개발



## 빅 데이터 시대 준비

빅 데이터 시대를 맞이하여 준바할 사항들은 데이터 경제 시대를 대비하는 ***연결과 협력, 창의적 인력의 양성, 데이터 신뢰 환경의 구축*** 등으로 요약될 수 있다.

### *데이터 경제 시대를 대비하는 연결과 협력*

o 기업은 다가올 데이터 경제 시대를 이해하고 정보 고립 상태를 경계해야 성공할 수 있음.

o 데이터는 무한한 자원이기는 하나, 활용 가능한 자원의 영역은 상호 연결과 협력으로 더욱 확장될 수 있음

o 데이터 경제 시대에는 플랫폼, 오픈 소스, 초고속 컴퓨팅 파워의 영향력이 커지며, 상호 연결과 협력이 핵심 전략으로 부각됨. 단절된 정보가 제한적으로 활용되는 것을 막고, 사회 전반적, 통합적 데이터 수집 및 활용을 위한 협의와 공동의 참여가 필요함

o 국가적으로 유용한 데이터 자원을 창출하려면 공공 부문과 민간 부문이 통합되는 데이터 수집, 분석 플랫폼 등 기반이 조성되어야 함. 이것은 참여자 간 신뢰, 상호 연결성, 공동 활용 방안 마련 등이 우선되어야 상호 합의로 발전 가능함

### *빅 데이터 핵심 역량인 창의적 인력의 양성*

o 빅 데이터는 수많은 데이터를 수집, 축적하는 것보다 무엇을 분석할 것인지 분명한 목적의식과 통합적 사고력, 해석력이 중요함

o 빅 데이터 시대를 맞이하여 미국에서는 2018년이 되면 14~19만 명의 분석전문가와 150만 명 정도의 데이터 관리자 등 분석 인력이 부족해질 것으로 전망함

o 향후 5년 동안 데이터 과학자의 수요가 공급을 뛰어넘어 인재 부족은 더욱 심화될 것임

o 빅 데이터 처리 인프라를 개발하고 빅 데이터 분석 플랫폼을 구축하는 시스템 엔지니어, 무엇을 분석할 것인지 분명한 목적의식과 통합적 사고력, 해석력을 갖춘 데이터 과학자, 직관적인 통찰과 풍부한 정보를 동시에 제공하는 가시화 아티스트 등이 빅 데이터 시대가 요구하는 핵심 인력임

o 빅 데이터 전문가에게는 대규모 데이터를 분석한 결과를 시각화하여 이해하기 쉽게 전달하는 역량이 필요함. 특히 시각화는 데이터 분석 결과를 전달하는 마지막 단계로서 데이터의 해석 작업에 해당되며, 정교한 모형 및 시각화 도구가 요구됨

o 국내외적으로 빅 데이터 인력의 부족을 인식하고, 유능한 인재 육성 프로그램과 교육 과정의 개설이 필요함
  - 단기간에 부족한 인력을 양성하고 더욱 빠르게 기술을 발전시키려면 빅 데이터를 기반으로 산학연이 협력하여 해결하는 국가적인 중대형 프로젝트가 절실히 요구됨
  - 글로벌 정보 기술 업체들도 데이터 과학자 확보에 심혈을 기울이며, 인재 확보와 내부 역량 강화에 노력함
  - 이베이에서는 고객 데이터를 분석하고 가공하는 일을 맡은 직원만 500명에 이름
  - EMC는 경제학, 통계학, 심리학 등을 전공한 박사급 인재들인 데이터 과학자로 구성된 애널리틱스 랩을 운영하며, 미국 IBM은 사내에 200명 이상의 수학자들이 분석학을 집중적으로 연구하여 500개 이상의 관련 특허를 취득하면서 미래 사업을 준비하고 있음

***데이터 신뢰 환경의 구축***

o 소셜 미디어에 있는 메시지, 흔적이나 개인의 정보가 담긴 빅 데이터 분석은 프라이버시 침해의 위험이 있음

o 접속 기록, 검색 패턴, 데이터 속성이 기록된 그림자 데이터의 증가는 프라이버시 침해를 위협하고 소셜 네트워크 기술 발전은 개인의 사적 공간을 훼손하고 있음

o 하지만 장기적으로는 사적 공간의 경계가 점차 모호해지면서 분쟁이 늘어날 것임

o 개인의 프라이버시를 보호하는 문제는 개인 정보 제공자나 개인 정보 활용자 모두에게 매우 중요한 품제임. 데이터에 민감한 개인 사용자 정보의 노출 없이도 타당한 수준의 분석을 도출할 수 있는 방안을 고려해야함.



# 참고 문헌

Bogdan Nedelcu, 2013, About Big Data and its Challenges and Benefits in Manufacturing, Database System Journal

David Corrigan, 2013, Integration and governing big data, IBM Software White Paper

Dunren Che, Mejdl Safran, and Zhiyong Peng, 2012, From Big Data to Big Data Mining: Challenges, Issues, and Opportunities, Database System for Advanced Applications

HITACHI Japan Homepage, http://www.hitachi.co.jp/products/it/bigdata/index.html

Julian Krumeich, Dirk Werth, Peter Loos, and Sven Jacobi, 2014, IEEE International Congress on Big Data

McKinsey&Company, 2012, Manufacturing the future: The next era of global growth and innovation

Shani, A.B., Sena, James A., 2000, Knowledge Management and New Product Development: Learning from a Software Development Firm

Tushar Rajpathak and Atul Narsingpurkar, 2015, Managing Knowledge from Big Data Analytics in Product Development

Usama M. Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth, 1995, From Data Mining to Knowledge Discovery: An Overview

일본 총무성, 2012, 정보통신백서(제2장, 스마트혁명이 가져오는 ICT산업ㆍ사회의 변혁)

일본 총무성, 2012, 빅 데이터 활용방안에 대하여

정지선, 2011, 새로운 미래를 여는 빅 데이터 시대, 한국정보화진흥원

# 부록: 빅 데이터 중요 기술 요약

## 빅 데이터에 관련 중요 기술

빅 데이터 활용에 관련된 중요 기술을 정리하면 다음과 같다.

▷ **NoSQL(Not only SQL(Structured Query Language)**

- o 표 형식으로 관계형 데이터베이스관리시스템(RDBMS)과는 다른 설계로 구현된 데이터베이스 시스템
- o RDBMS가 정형 데이터 처리를 필요로 하는 업무시스템의 이용에 적합한 반면
- o NoSQL은 센서 및 소셜 미디어 등 비정형 데이터를 포함한 다양한 데이터를 대량으로 데이터베이스화에 사용

▷ **Hadoop**

- o 미국 NPO의 Apache 소프트웨어재단의 프로젝트로 개발이 진행되고 있는 대규모 데이터의 효율적인 분산 처리 등을 위한 오픈 소스 소프트웨어 프레임워크
- o 여러 서버를 통한 병렬 처리에 의해 유연하고 지속적인 대규모 데이터의 고속 처리가 가능

▷ **클라우드 서비스**

- o 이용자가 필요한 컴퓨터 자원을 "필요한 때에 필요한 양만" 이용이 가능하여 확장성, 가용성, 민첩성 및 경제성 등의 특징을 갖춘 서비스
- o 클라우드 서비스를 이용하면 다종 다량의 데이터의 축적 및 계산 처리를 위해 필요한 여러 시스템을 자기 부담으로 준비 할 필요없이 저렴한 비용으로 비슷한 환경의 구축이 가능
- o 예를 들면 Hadoop 실행 환경을 제공하는 서비스를 이용하면 시스템의 구매뿐만 아니라 소프트웨어 설치 등의 설정 작업의 생략도 가능

▷ **DWH(Data Ware House)**

- o 정형 데이터·비정형 데이터를 불문하고 대량 데이터의 축적을 목적으로 하는 데이터베이스의 총칭
- o 여기에는 대량의 데이터를 고속으로 처리 하는 방법에 따라 다음 2가지 방식으로 분류함
  - ① 전통적인 RDBMS와 다른 방식에 의한 NoSQL 데이터베이스
  - ② 표준 RDBMS와 하드웨어 수준의 속도 향상 기술을 결합한 DWH 어플라이언스

### ▹ CEP(Complex Event Processing)

- o 데이터를 디스크에 저장하지 않고 쓰기 속도가 디스크에 비해 빠른 메모리에서 순차적으로 처리하여 필요한 정보를 실시간으로 추출하는 기술
- o 디스크에 데이터를 축적하고 분석하는 기법에 비해 단시간에 처리가 가능하기 때문에 신용 카드 부정 사용이나 방범 카메라 영상의 이상 검지 등 단기간에의 대응이 필요한 경우에 사용
- o 미리 이용자가 정의하는 실시간 처리의 내용에는 단일 데이터 속성의 임계값에 의한 판별 외에도 여러 특성을 조합한 처리 설정도 가능

### ▹ PPDM (Privacy Preserving Data Mining)

- o 개인 정보를 보호하면서 데이터에서 특징이나 규칙성 등을 추출하는 기술
- o 익명화와 은밀한 계산 등에 따라 개별 데이터를 암호화한 채로 데이터 마이닝을 실시하여 개인 정보 유출 등의 위험을 방지하면서 데이터 분석 등이 가능
- o 예를 들면 어떤 데이터에 대해서도 같은 것이 k 개 이상 존재하도록 데이터의 입도 및 모호함을 제어하는 k-익명화 기술 등이 있음

### ▹ MDM (Master Data Management)

- o 업무상 가장 기본적인 정보인 고객 정보 등의 마스터 데이터를 관리하기 위한 시스템
- o 다양한 정보 시스템에 중복·산재하며 다량으로 생성되는 마스터 데이터에 대해 항상 최신 상태로 업데이트하고 전체 시스템의 무결성 확보 등을 지원

### ▹ 비밀 계산

- o 입력 데이터 및 연산 논리를 암호화된 상태로 모든 계산을 가능하게 하는 기술
- o 여러 컴퓨터에 데이터 조각을 보내고 단편 부분 계산을 반복함으로써 데이터를 은닉 한 채로 통계 등 각종 계산이 가능
- o 민간 기업, 공공 기관, 교육 현장 등의 프로그램의 잘못된 해석 방지, 지적 재산권 침해 방지, 정보 유출 방지 등 다양한 분야에 응용이 가능



빅 데이터 활용에 빅 데이터 관련 기술의 활용을 나타내면 다음과 같다:

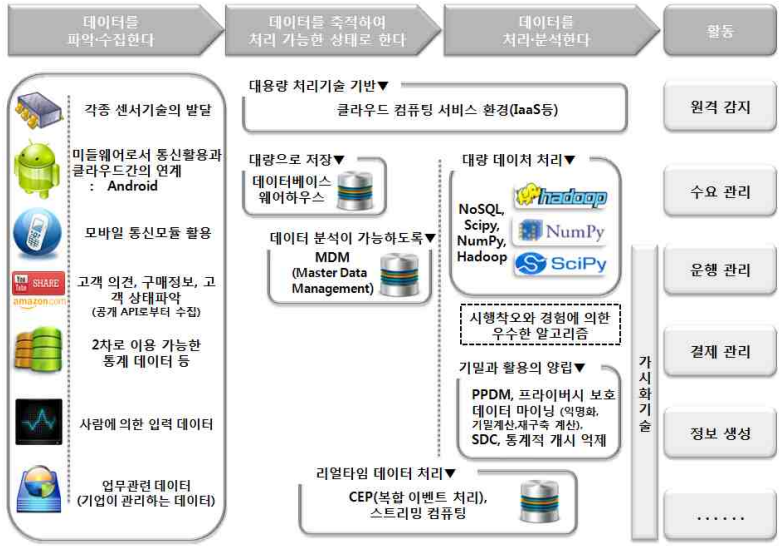



그림 47. 빅 데이터 활용 및 기술 적용 이미지