빅 데이터와 공공 정책

전승우*
LG경제연구원 선임연구원
swjeon@lgeri.com

1. 빅 데이터에 대한 뜨거운 관심

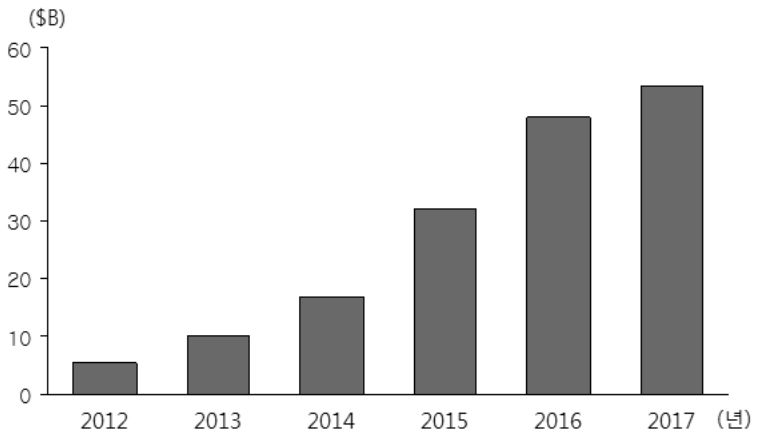
산출되는 데이터의 규모가 하루가 다르게 증가하면서 빅 데이터는 IT를 넘어 다양한 분야의 뜨거운 화두로 주목받고 있다. 전세계적으로 매일 2.5 엑사바이트(Exabytes)의 엄청난 데이터가 새롭게 생성되고 있으며, 그 산출 속도 또한 지속적으로 증가하고 있다고 한다. 특히 트위터와 페이스북 등 소셜 네트워크의 데이터가 급속하게 늘고 있는 가운데, 스마트폰과 태블릿 PC를 비롯한 모바일 기기들의 보급이 확산되면서 이동통신 역시 새로운 데이터 창출의 원동력으로 급부상하고 있다.[1)]

따라서 많은 리서치 및 컨설팅 기업들은 폭증하는 데이터의 규모와 잠재력에 대한 전망을 발표하고 있다. IT 전문 리서치 기관 가트너(Gartner)는 다양한 분야에 걸쳐 무한한 가치를 지니고 있는 빅 데이터가 21세기의 원유라 될 것이라고 주장하였다. 또한 글로벌 컨설팅 기업 맥킨지(Mckinsey)는 빅 데이터의 활용을 통하여 의료, 공공행정, 소매, 제조, 개인정보 등의 부문에서 1%의 생산성을 추가로 향상시킬 수 있으며, 각 분야별로 최

* 저자 학력, 경력 및 최근 연구:
연세대학교 기계전자공학부 학사, 한국과학기술원 전자전산학 석사, 서울대학교 경영전문대학원 MBA
전) 삼성종합기술원 Future IT Lab 연구원, 현) LG경제연구원 사업전략부문 선임 컨설턴트
플랫폼 경쟁 이어 모바일 AP 경쟁 치열해지고 있다(2013), 특허전쟁시대, 특허전문기업의 화력 강해지고 있다(2012), 빅 데이터에 대한 기대와 현실(2012), 모바일 트래픽 폭증시대 네트워크 품질의 중요성 커진다(2012)

1) "Big data: The management revolution", Harvard Business Review, Oct. 2012

소 1,000억 달러에서 최대 7,000억 달러 규모의 경제적 효과를 창출할 수 있다는 전망을 내놓았다.[2)]



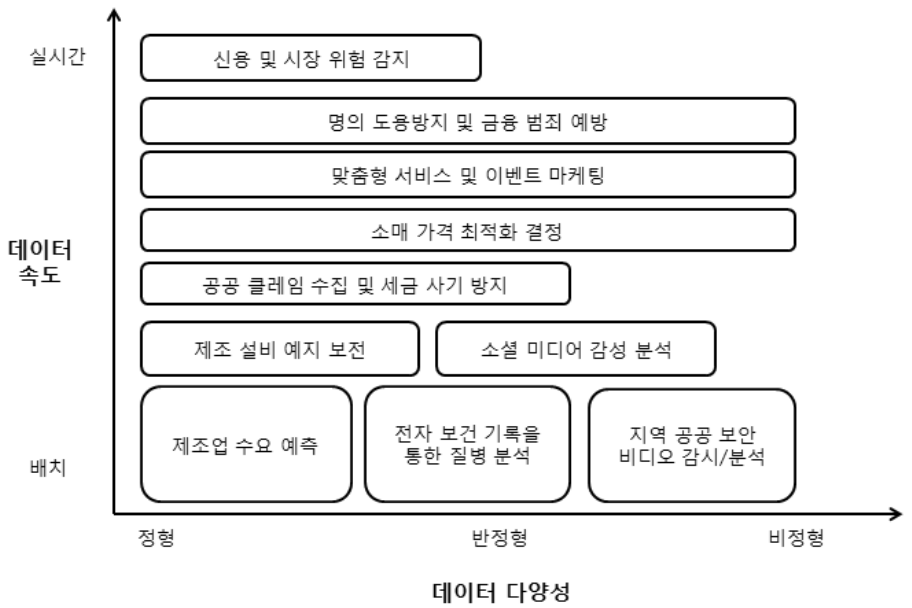
자료: Wikibon

[그림 1] **글로벌 빅 데이터 시장 규모**

아직까지 각 기관마다 빅 데이터를 바라보는 시각은 조금씩 상이하지만, 이들은 공통적으로 빅 데이터가 새로운 현상 및 미래 트렌드를 정확히 파악하고 이전에 없던 새로운 가치를 창출할 수 있는 자원이 될 것이라고 말한다. 특히 지금까지는 일정한 형식과 규칙에 따라 데이터베이스에 저장되어 있는 정형 데이터의 활용이 주를 이루었던 반면, 향후 빅 데이터 시대에서는 전체 데이터의 80% 이상을 차지하고 있는 음성, 영상, 텍스트 등 정형화되지 않은 데이터가 중요한 전략적 자산이 될 것이라 설명한다.

이처럼 빅 데이터의 가치에 대한 공감대가 확산되고 구글(Google)과 아마존(Amazon), IBM 등 여러 IT 기업을 중심으로 빅 데이터 활용 사례가 알려지기 시작하면서, 빅 데이터는 여러 산업에 걸쳐 큰 주목을 받게 되었다. 구글은 인터넷을 통하여 수집한 엄청난 규모의 문서를 이용하여 다양한 언어로 실시간으로 번역할 수 있는 자동 번역 시스템을 개발하였으며, 아마존은 고객의 도서 구매 데이터를 분석하여 향후 추가로 구매할 것으로 예상되는 도서를 추천하고 할인 쿠폰을 지급하는 마케팅 활동으로 큰 호응을 얻었다. 또한 세계 최대의 유통 기업 월마트(Walmart)는 고객으로부터 엄청난 규모의 데이터를 매 시간마다 실시간으로 수집하고 분석하고 있으며, 글로벌 패스트패션 기업 자라(Zara)는 전세계의 고객 구입 및 재고 데이터를 통하여 최신 유행 트렌드를 빠르게 감지하고 있다.

2) "Big data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute, May 2011



자료: IDC

[그림 2] **빅 데이터 분석의 활용 사례**

2. 빅 데이터의 공공 활용 사례

빅 데이터가 아직까지는 민간을 중심으로 활성화되고 있지만 여러 선진국 역시 빅 데이터에 대한 연구 및 적용에 적극적으로 나서고 있다. 각국 정부는 이미 사회 각 분야에 걸쳐 다양한 종류의 데이터를 광범위하게 수집하고 있기 때문에 이를 기반으로 효과적인 정책을 수립하고 여러 현안들을 효과적으로 해결할 수 있을 것으로 기대하고 있다.

오늘날 IT 분야에서 확고한 경쟁력을 보유하고 있는 미국은 빅 데이터의 공공 정책 활용에 가장 앞선 움직임을 보이고 있는 것으로 평가된다. 백악관 과학기술정책실(OSTP)은 국립과학재단(NSF), 국방부(DOD) 등 6개 연방정부기관과 함께 'Big Data Research and Development Initiative'를 구성하여 빅 데이터의 기술 연구 및 활용 전략의 수립에 2억 달러를 투자하기로 발표하였다. 또한 대통령 과학기술자문위원회(PCAST)는 각 정부기관들이 향후 폭발적으로 증가하게 될 데이터의 새로운 가치에 주목하고 이에 기반을 둔 새로운 지식 창출 및 전략 수립에 적극적으로 나서야 한다고 강조하기도 하였다.

이에 따라 미국은 빅 데이터를 기반으로 다양한 공공 정책을 추진하고 이를 통하여 상당한 성과를 거두고 있다. 미국 국세청(IRS)은 사기 행위 분석, 소셜 네트워크 분석을 통한 범죄 네트워크 발굴, 데이터 분석을 통한 지능형 사기 감시 등 다양한 기술을 활용

하여 탈세 및 사기범죄예방 시스템을 구축한 결과, 세금 누락 및 불필요한 세금 환급 절감을 통하여 연간 3,450억 달러를 절약할 수 있게 되었다. 한편 미국 국립 의료원(NIH)은 사용자가 요구하는 다양한 약에 대한 정보를 제공하고 제약사와 사용자간의 유기적인 상호 소통을 지원하는 필박스(Pillbox) 서비스를 통하여 수천 만 달러의 비용을 절감할 수 있었으며, 유행하는 질병의 발생 장소 및 전염 속도에 대한 연구에도 활용하고 있다.[3)]

특히 미국 본토를 강타한 9.11 테러 이후 애국법이 제정되는 등 철저한 사회 안보가 강조되면서 빅 데이터는 각종 치안 유지 및 보안 활동에도 활발하게 적용되고 있다. 뉴욕 경찰국(NYPD)이 운영하고 있는 실시간 범죄정보센터는 IBM의 빅 데이터 분석 기술을 이용하여 범죄 기록과 판결문, 체포 기록 등 수백 만 건의 데이터를 경찰관에게 실시간으로 제공하고 있다. 또한 미국 중앙정보국(CIA)는 세계 각지의 소속 인원들이 수집하고 있는 데이터를 기반으로 각국의 정세 변화 및 테러 가능성을 실시간으로 예측하고 있으며, Digital Reasoning, Recorded Future, Palantir Technologies 등 여러 데이터 분석 기업과의 밀접한 협력을 통하여 빅 데이터 활용 역량을 체계적으로 강화하고 있다. 한편 빅 데이터의 활용에 적극적인 싱가포르 정부 역시 2004년부터 빅 데이터 기반 위험 관리 계획을 수립하고 국가의 위험을 평가하고 다양한 정세 변화를 탐지하는 국가위험관리시스템(Risk Assessment Horizon Scanning)을 운영하고 있다.

3. 빅 데이터를 둘러 싼 논란

1) 빅 데이터, 또 하나의 IT 거품일까?

빅 데이터에 대한 사회경제적 관심이 뜨겁게 고조되고 있음에도 불구하고 빅 데이터의 개념은 아직 모호한 수준이다. 특히 빅 데이터가 단지 거대한 데이터만을 지칭하는 것이 아니라 이를 처리하기 위한 기술과 시스템, 나아가서는 활용 전략 등 다양한 의미를 내포하고 있기 때문에, 이를 체계적으로 이해하기란 그리 쉽지 않다. 무엇보다도 빅 데이터가 기존의 IT 기술에 비해 어떻게 차별화된 가치를 창출할 수 있을 것인지에 대한 의견이 분분하고, 한편으로는 빅 데이터가 결국 CRM, ERP, 웹서비스와 같이 잠시 세간의 이목을 집중시키는 IT 유행어(Buzzword)에 그칠 것이라는 회의론도 제기되고 있다. 특히 빅 데이터에 대한 논의가 아직까지는 대용량 데이터를 신속하고 저렴하게 처리할 수 있는 Hadoop과 NoSQL 등 주요 기술 및 시스템에 주로 국한되어 있으며 일부에서는 통상적인 데이터 분석이 빅 데이터의 대표적인 성공 사례로 소개되기도 하는 등, 빅 데이터에

3) “빅 데이터 글로벌 10대 성공사례”, 한국정보화진흥원, 2012.04

대한 올바른 인식과 접근이 이루어지기까지는 보다 오랜 시간이 소요될 것이라는 의견도 설득력을 얻고 있다.

사실 IT에 대한 투자가 생산성과 비례하지 않는다는 생산성 패러독스(Productivity paradox)의 주장이 제기된 이래 IT의 성과 창출 효과에 대해서는 여전히 꾸준한 논쟁이 벌어지고 있다. 따라서 빅 데이터 역시 뜨거운 관심이 이어지고 있음에도 불구하고 그 효용성에 대한 논란에서 자유로울 수 없을 것으로 전망된다. 더군다나 빅 데이터에 대한 체계적인 연구 및 활용이 아직 일천한 가운데, 오픈 소스 및 클라우드 컴퓨팅 등 IT 자원의 빠른 가격 하락에도 불구하고 빅 데이터의 수집과 분석은 시간과 직/간접 비용 등 다각적인 측면에서 여전히 부담이 큰 투자이다. 그러므로 가트너가 빅 데이터에 투자한 기업의 85% 이상이 성과를 거두지 못할 것이라는 비관적인 전망을 발표하기도 하는 등, 일부를 제외하고는 빅 데이터에 대한 대부분의 투자가 만족할 만한 성과를 거두지 못할 것이라는 예상도 이어지고 있다.

2) 빅 데이터, 보안을 위협하다

오늘날 빅 데이터에 대하여 가장 많이 제기되고 있는 우려는 바로 민감한 데이터의 누출 및 개인의 사생활 침해 등 각종 보안과 관련된 문제이다. 다양한 경로를 통하여 개인의 데이터를 수집하고 축적하는 과정에서 민감한 사적 정보들이 손쉽게 유출될 가능성이 점점 증가하고 있다. 일상 전반에 걸쳐 인터넷의 사용이 확산되고 해킹 등 악의적인 기술이 이를 저지하는 보안 기술보다 빠르게 발전하면서, 이제 누구라도 마음만 먹으면 타인의 중요한 데이터를 손쉽게 획득할 수 있는 환경이 조성되고 있다. 따라서 빅 데이터가 본격적으로 언급되기 이전에도 데이터 보안과 관련된 문제는 이미 민간과 공공을 막론하고 끊임없이 제기되어 왔다.

게다가 기업과 정부의 광범위한 데이터 수집 활동은 그 의도를 떠나 대중을 감시하고 통제하는 '빅 브라더(Big brother)'가 될 수 있다는 경각심을 부각시키면서 큰 반발과 저항을 야기할 위험이 크다. 세계 최대의 인터넷 기업 구글이 이메일, 지도, 동영상 등 60여 개의 서비스를 이용하는 10억 명의 개인정보를 모두 통합해서 관리하는 개인정보 통합관리를 실행하겠다고 밝히면서 논란의 중심에 서 있는 가운데, 애플(Apple)은 아이폰을 사용하는 고객들의 위치 정보를 무단으로 수집하여 미국에서 150억 달러 규모의 집단 소송에 휘말리기도 하였다. 또한 사진 속 인물을 자동으로 인식해 이름을 태그하는 얼굴 인식 서비스를 수행하던 페이스북(Facebook)은 유럽 사용자들의 거센 반발에 부딪히자 유럽 지역의 서비스를 제한하고 이미 수집한 사용자들의 얼굴 정보를 모두 삭제하는 소동을 겪었다. 무엇보다도 최근 미국 국가안보국(NSA)과 연방수사국(FBI)이 2007년

부터 마이크로소프트(Microsoft)와 구글, 페이스북, 애플 등 9개 기업의 고객 정보를 수집하는 프리즘(Prism) 프로그램을 추진해 온 것이 드러나면서 국민들에게 큰 충격을 안겨 주었다. 특히 국가안보국은 주요 국가의 대사관을 감청하여 기밀 정보를 수집해 온 것으로도 드러나 이러한 빅 브라더 논란은 전세계적으로 확대될 움직임을 보이고 있다.

3) 빅 데이터 분석의 가능성과 한계

많은 전문가들은 거대한 데이터 분석을 통하여 현재의 현상 및 향후 발생 가능한 사건들을 보다 정확히 예측할 수 있게 될 것이라고 설명한다. IT 매거진 와이어드(Wired)의 편집장 크리스 앤더슨은 수많은 데이터간의 연관성을 분석함으로써 이론적 모델 없이도 새로운 사실을 발견하고 이를 기반으로 각종 문제를 해결할 수 있을 것이라고 주장하기도 하였다.

엄청난 양의 데이터를 통계적으로 분석하고 이를 기반으로 의사 결정을 수행하는 것은 주먹구구식의 즉흥적인 판단으로 야기할 수 있는 편향적 오류를 낮추는 데에 큰 도움이 될 수 있다. 실제로 많은 기업 및 정부는 미래에 발생할 수 있는 여러 시나리오를 다양한 데이터를 기반으로 정교하게 평가함으로써 가장 합리적인 전략을 수립하고자 노력하고 있다.

그러나 이러한 빅 데이터 분석이 모든 현상을 정확히 판단하고 예측할 수 없다는 경계론도 고개를 들고 있다. 근본적으로 과거의 정보를 담고 있는 데이터는 현재 및 미래와의 상관관계가 높지 않을 수 있으며, 데이터의 출처에 따른 편향성 및 정보 부정확성, 혹은 데이터 분석이 사실을 과장 혹은 축소 해석할 위험 등 다양한 통계적 오류가 여전히 발생할 수 있다는 것이다. 실제로 2008년 발발한 금융 위기 당시 미국의 유수 투자은행들은 IT 시스템에 10년 간 수백 억 달러를 쏟아 부었지만, 서서히 다가오는 서브프라임 모기지의 위험에 전혀 대처하지 못하고 하루아침에 무너져 버리고 말았다.

또한 오늘날 빅 데이터의 대표적인 사례로 부각되고 있는 소셜 네트워크 등 비정형 데이터에 대한 분석 역시 실제보다 그 성과가 과장되었다는 주장도 만만치 않다. 물론 비정형 데이터에 대한 분석 기술이 빠르게 발전함에 따라 이전에 알지 못하였던 다양한 사실과 인과 관계를 발견하는 데에 큰 역할을 하고 있다. 실제로 미국 인디애나 주립대의 페비오 라저스 교수는 2010년 미국 연방하원의원 선거 기간 동안 모든 선거구에서 트위터에 기반한 후보 예측이 93%의 적중률을 기록했다는 연구 결과를 발표하기도 하였다.[4] 그러나 비정형 데이터의 출처 및 종류와 특성이 각기 천차만별이므로 이러한 데이터의 수집 및 가공 과정에서 다양한 오류가 발생하고 실제 결과의 가치도 분석의 노력에 비해

4) "How Twitter can help predict an election", Washington Post, Aug. 2013

크지 않다는 높다는 지적도 제기되고 있다. 일례로 소셜 네트워크는 도시의 젊은 계층이 주로 이용하고 있으며 대부분의 데이터가 객관적인 사실보다는 개인의 주관적인 성향을 더욱 많이 포함하고 있으므로 정확한 트렌드와 여론을 구체적으로 파악하기 어려우며, 나아가 이를 기반으로 자칫 잘못된 예측을 하게 될 위험도 클 수 있다.

4. 빅 데이터의 공공 정책 활용 전략

1) 철저한 빅 데이터의 활용 목적과 실행 방안 마련

모든 분야에 공통적으로 적용될 수 있는 빅 데이터 전략이란 존재하지 않는다. 동일한 수준의 데이터 분석이라 하더라도 각 기업 및 정부의 데이터 활용 목적과 주요 대상 및 주변 환경에 따라 서로 다른 결과와 해석을 얻을 수 있기 때문이다.

특히 공공 정책은 기업의 경영 전략과 달리 다양한 사회 구성원을 대상으로 각종 현안을 다룬다는 점에서, 각각의 정책마다 필요한 빅 데이터의 특성 및 활용에 따른 효과는 매우 상이할 것으로 보인다. 따라서 뚜렷한 방향성이 결여된 빅 데이터의 도입은 의도와 달리 적극적으로 활용되지 못하거나 오히려 과도한 투자로 인하여 정책 추진 상 비효율성을 야기하게 될 수 있다.

그러므로 정부는 각 정책에 따른 빅 데이터의 수집과 분석 범위를 정의하고 이를 어떻게 활용할 수 있을지를 사전에 명확히 설정하는 것이 중요하다. 즉 해당 정책 분야의 고유한 특성을 바탕으로 데이터 분석의 목적과 이를 통한 기대 효과의 청사진을 제시하고 이를 위하여 필요한 시스템 및 기술의 도입, 분석 결과의 활용과 가치 창출 방안 등 구체적인 실행 프로세스를 마련해야 한다. 아울러 이러한 전략이 원활하게 수행될 수 있도록 빅 데이터에 대한 내부 구성원들의 충분한 공감대 형성에 주력하는 동시에, 데이터 분석 결과 적용의 한계와 그 대안에 대한 고려도 마련해야 할 것이다.

2) 데이터의 질적 수준 강화 노력

충분한 규모의 데이터가 분석의 정확성 및 신뢰도를 높이는 것이 사실이나, 데이터의 절대적 규모만이 기대한 결과를 효과적으로 얻을 수 있는 필수 요건은 아니다. 거대한 양의 데이터와 이를 정교하게 처리할 수 있는 최신 분석 기술을 가지고 있더라도, 데이터 자체의 질적 수준이 기대에 미치지 못한다면 분석에 따른 시간과 비용이 증가할 뿐 자칫 원하는 결과를 얻는 데에 실패할 수도 높다. 따라서 수집한 데이터가 정확한 정보를 내포하고 있는지, 그리고 분석 목적에 맞게 유용하게 활용될 수 있는지 등 데이터의

질적 수준에 대한 충분한 판단이 이루어져야 한다.

실제로 많은 연구에서는 데이터의 양보다는 질적 수준이 가치 있는 지식과 통찰력 제공에 기여한다고 주장하고 있다. 두 차례의 미국 대통령 선거 결과를 정확하게 예측하여 유명 인사로 떠오른 데이터 과학자 네이트 실버 역시 데이터의 규모보다는 필요한 데이터를 정확하게 분별하고 수집하는 능력이 더욱 중요하다고 지적하기도 하였다. 아무리 IT 기술이 발전해도 데이터의 질적 수준을 판별하는 것은 결국 유능한 인재의 몫으로 남게 될 것이다. 따라서 문제 해결에 도움을 줄 수 있는 올바른 데이터를 수집하기 위한 조직의 전문성과 창의력이 빅 데이터 시대의 중요한 과제로 떠오를 것으로 전망된다.

3) 빅 데이터의 효용과 안전의 균형 유지

마지막으로 빅 데이터의 기회 및 잠재력과 더불어 향후 급격히 확산될 보안 사고의 위험을 인지하고 이에 대한 만반의 준비를 갖추는 것도 필요하다. 빅 데이터의 누출과 감시 등 보안과 관련된 문제는 평소에 쉽게 드러나지 않더라도 일단 발생하면 빅 데이터의 장점을 한 순간에 무너뜨릴 수 있는 파괴력을 지니고 있다. 따라서 최근 불거지고 있는 사이버 테러 및 기업의 고객 정보 유출, 그리고 시스템의 복잡화 및 과부하에 따른 네트워크 마비 사태 등은 빅 데이터 시대에 더욱 큰 위험으로 다가올 전망이다.

게다가 잘못된 빅 데이터 활용은 정부의 데이터 수집에 대한 논란과 거부감을 확대시킬 위험이 크다. 정부가 개인의 세밀한 정보를 담고 있는 데이터에 대한 접근성이 높다는 점은 빅 데이터 활용의 효과를 높이는 데에 도움을 줄 수 있다. 그러나 한편으로 데이터를 통한 개인의 사생활 침해 등 악용의 사례가 발견될 경우 정부에 대한 불신과 반발 등 더욱 큰 역풍을 맞게 될 수 있다는 점을 유념할 필요가 있다.

물론 빅 데이터의 활성화를 위해서는 이전보다 적극적인 정보 수집 및 공개적 활용이 불가피하다는 의견도 적지 않다. 그러나 최근 미국의 빅 브라더 논란에서 알 수 있듯이, 개인 및 공공의 정보가 사회적 합의 없이 유출되거나 수집될 경우에는 걷잡을 수 없이 큰 혼란을 야기하게 될 것으로 예상된다. 따라서 데이터의 효용과 안전 사이에서 적절한 균형을 유지하기 위한 정부의 세심한 빅 데이터 정책 수립 및 추진 노력이 요구된다. 데이터 누출을 차단할 수 있는 시스템의 촘촘한 구축은 물론이고 데이터 활용 실태 점검과 가이드라인의 마련, 인력에 대한 교육 등 지속적인 보안 강화의 노력만이 빅 데이터로 촉발될 수 있는 엄청난 위험을 최소화할 수 있을 것이다.