

빅 데이터 비즈니스의 이슈와 전망

류 한 석 (Ryu, Hanseok)

- 現류한석기술문화연구소장, IT 칼럼니스트

- 블로그 <http://peopleware.kr> / 트위터 @bobbyryu

1. 빅 데이터의 등장 배경 및 특징

작년 IT업계의 핫 이슈가 '클라우드(Cloud)'였다면 올해는 '빅 데이터(Big Data)'가 아닐까? IT업계는 주기적으로 새로운 이슈를 만들어낸다. 마치 그런 것만 고민하는 사람들이 있는 것 같다. 클라우드가 과거의 ASP(Application Service Provider)나 SaaS(Software as a Service)의 연장선상에 있듯이, 빅 데이터는 데이터 웨어하우스(Data Warehouse)나 비즈니스 인텔리전스(Business Intelligence)의 연장선상에 있다고 볼 수 있다. 전혀 새로운 기술은 아닌 것이다. 먼저 빅 데이터의 등장 배경을 살펴보자.

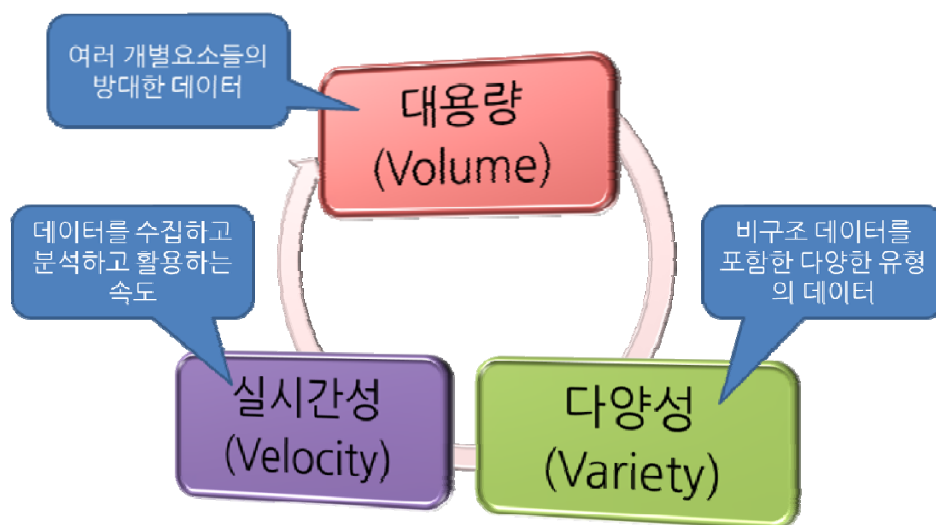
현재 우리나라를 비롯한 대부분의 선진국들에서는 업무자동화를 위한 정보시스템의 구축이 거의 완료된 상태다. 기존 업무 프로세스에 수정이 발생하거나 또는 시스템을 업그레이드하거나, 아니면 새로운 비즈니스에 진출함으로써 생기는 신규 정보시스템 수요가 있기는 하지만 덩치가 커진 IT서비스 기업들의 입장에서는 양에 차지 않는 실정이다. 뿐만 아니라 클라우드의 영향으로 인해 전체 IT 시장은 축소되는 경향을 보이고 있으며 앞으로 더욱 더 가속화될 전망이다. 또한 기업들이 쌓아두는 정보는 계속 방대해지고 있는 반면에 정보를 분석하고 예측하는 일은 점점 어려워지고 있다.

이와 같은 배경으로 인해 IBM, HP, 삼성 SDS, LG CNS 등과 같은 IT서비스 기업들은 새로운 부가가치의 창출이 절실해졌으며 그에 따라 클라우드 시대에 걸맞은 새로운 수익원으로서 빅 데이터 비즈니스를 주목하게 됐다. 또한 일반 기업들(IT서비스 기업의 입장에서는 고객사)은 이제는 단순 정보화가 아니라 자사가 확보한 데이터를 통해 비즈니스 분석/예측 역량을 강화함으로써 점차 치열해지고 있는 시장경쟁에서 앞서 나가야겠다는 생각을 갖게 됐다.

빅 데이터는 말 그대로 막대한 양의 데이터다. 빅 데이터 기술에는 1) 다양한 형태의 데이터를 수집하고 통합하는 것, 2) 데이터를 분석해 트렌드와 패턴을 찾아내는 것, 3) 분석 결과를 실시간으로 활용해 비즈니스 향상에 기여하는 것 등의 여러 단계가 모두 포함된다.

빅 데이터의 3대 특징을 정리해보면 다음과 같다.

- ① 대용량(Volume): 과거보다 데이터의 규모가 더욱 증가했다. 이는 여러 개별요소들의 방대한 생데이터(Raw Data: Source Data 또는 Atomic Data라고도 함)의 집합이다.
- ② 다양성(Variety): 빅 데이터에서는 기존의 관계형 데이터베이스뿐만 아니라 SNS, 위치정보, 각종 로그 기록을 비롯해 멀티미디어 등의 비정형 데이터를 포함한 다양한 유형의 구조화되지 않은 데이터를 다룬다.
- ③ 실시간성(Velocity): 데이터를 생성하거나 수집 및 통합하고 분석하고 활용하는 모든 단계에 있어서 속도가 중요하다. 궁극적으로 빅 데이터에서는 분석 결과를 실시간으로 활용하는 것을 추구하며, 이것이야말로 과거의 유사한 기술 트렌드와 빅 데이터를 구별하는 가장 큰 특징이라 할 수 있다.



[그림] 빅 데이터의 3대 특징

인터넷에서 발생하는 데이터의 양은 꾸준히 증가해왔는데 최근에는 SNS와 스마트폰의 대중화로 인해 데이터가 급증하고 있는 추세다. 또한 정부기관과 개별 기업들은 정보시스템을 통해 방대한 데이터를 확보할 수 있게 되었으며, 스스로 확보한 데이터뿐만 아니라 타사의 데이터 또는 인터넷 상의 데이터를 통합하고 분석하여 비즈니스 인사이트(Insight)를 찾아내고 이를 정책 또는 비즈니스 향상에 활용하려는 욕구가 점차 커지고 있는 상황이다.

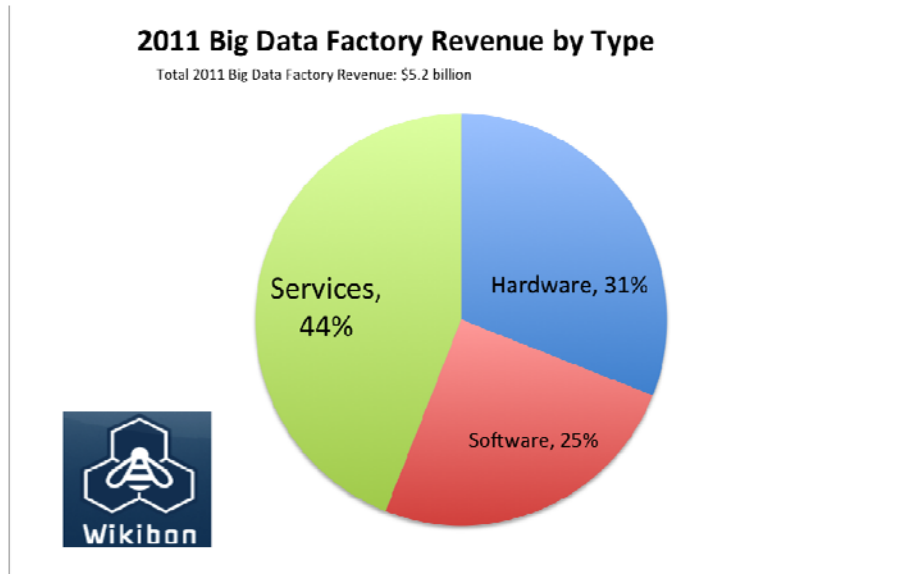
컨설팅업체 맥킨지는 빅 데이터를 제대로 활용하면 공공, 제조, 소매, 의료 부문에서 1%의 생산성을 추가로 향상시킬 수 있다고 밝혔다. 또한 EU(유럽연합)는 연간 2500유로 이상, 한국은 10조 7000억 원 이상의 정부 지출을 줄일 수 있을 것이라고 전망했다.

미국의 경우 국가 차원에서 빅 데이터 기술 개발을 적극 지원하고 있다. 미대통령 직속 과학기술정책실은 'Big Data Research and Development Initiative'에 2억 달러를 투입한다고 발표했으며, 지식과 인사이트를 추출해내는 능력을 향상시키는 것이 목적이라고 밝혔다. 미국방부는 빅 데이터를 활용해 스스로 인지하고 결정해 군사 행동을 수행하는 자율 시스템을 개발할 계획이며, 이를 위해 상금을 걸고 콘테스트도 개최할 예정이다. 미 국립보건원은 AWS(Amazon Web Services)를 통해 세계 최대의 유전자 변형 데이터 세트(200TB의 용량)를 무료로 공개한다고 발표하기도 했다.

이처럼 빅 데이터가 개별 기업의 경쟁력뿐만 아니라 국가 경쟁력으로 부상함에 따라 미국, 중국, 일본, EU를 비롯한 전세계 각국은 빅 데이터 전략 수립에 적극적으로 나서고 있다.

2. 빅 데이터 기술 및 업체 동향

전문가집단인 위키본(Wikibon)에 따르면, 2011년 빅 데이터 시장 규모는 52억 달러였으며 하드웨어, 소프트웨어, 서비스 분야 중 서비스 매출이 44%로서 가장 큰 비중을 차지했다.





[그림] 2011년 빅 데이터 시장 규모 (출처: 위키본)

앞으로 빅 데이터 시장에서 앞서가기 위해서는 서비스 역량이 가장 중요할 것으로 판단된다. 왜냐하면 빅 데이터의 궁극적인 목적은 기술의 도입이 아니라 이를 통해 비즈니스 인사이트를 도출하고 실제로 비즈니스를 향상시키는 것이기 때문이다.

올바른 비즈니스 목표를 수립하고 그에 부합하는 트렌드와 패턴을 찾아내고 해당 내용을 어떻게 비즈니스에 활용할 것인가를 결정하고 실행하는 것이 바로 서비스 역량이다. 이를 위해 기업은 스스로 그런 역량을 갖추든가 아니면 IT서비스 기업의 도움을 받아야 한다. 현재 IT서비스 기업들은 하드웨어, 소프트웨어, 서비스 분야에서 각각의 강점을 바탕으로 빅 데이터 경쟁력을 강화하고 있으며 신생 업체들도 속속 등장하고 있다. 빅 데이터의 주요 분야 및 관련 기술/업체들의 목록은 다음과 같다.

BIG DATA Market Segments

Hardware	Big Data Distributions	Data Management Components	Analytics Layer	Applications Layer	Services
<ul style="list-style-type: none"> Storage Servers Networking <p>Vendors include Dell, HP, Arista, IBM, Cisco, EMC, NetApp.</p>	<ul style="list-style-type: none"> Open source Hadoop distributions Enterprise Hadoop distributions Non-Hadoop Big Data frameworks <p>Vendors/providers include Apache, Cloudera, Hortonworks, IBM, EMC, MapR, LexisNexis.</p>	<ul style="list-style-type: none"> Distributed file stores NoSQL databases Hadoop-optimized data warehousing Data integration Data quality and governance <p>Vendors/providers include Apache, DataStax, Pervasive Software, Couchbase, IBM, Oracle, Informatica, Syncsort, Talend.</p>	<ul style="list-style-type: none"> Analytic application development platforms Advanced analytics applications <p>Vendors/providers include Apache, Karmasphere, Hadapt, Attivio, 1010data, EMC, SAS Institute, Digital Reasoning, Revolution Analytics.</p>	<ul style="list-style-type: none"> Data visualization tools Business intelligence applications <p>Vendors include Datameer, ClickFox, Platfora, Tableau Software, Tresata, IBM, SAP, Microstrategy, Pentaho, QlikTech, Japersoft.</p>	<ul style="list-style-type: none"> Consulting Training Technical support Software maintenance Hardware maintenance Hosting/Big-Data-as-a-Service/cloud <p>Vendors include Tresata, Tidemark, Think Big Analytics, Amazon Web Services, Accenture, Cloudera, Hortonworks.</p>
  <p>Next Generation Data Warehouse Appliances</p>					
<ul style="list-style-type: none"> MPP, columnar data warehouse appliances. In-memory analytics engines Fast data loading <p>Vendors include EMC Greenplum, HP Vertica, Teradata Aster, IBM Netezza, Kognitio, ParAccel.</p>					

[그림] 빅 데이터 기술/업체 목록 (출처: 위키본)

빅 데이터는 클라우드 기반의 대용량 데이터 처리 기술인 하둡(Hadoop)을 비롯해, 전통적인 RDBMS(Relational Database Management System)를 보완하기 위한 NoSQL(Not only SQL), 그리고 각종 데이터 시각화(Data Visualization) 기법에 이르기까지 방대한 기술 세트를 사용한다. 여기에서 각각의 기술을 모두 설명할 수는 없으므로 구체적 내용이 궁금하다면 위의 표를 참고해 관련 기술을 검색해보길 추천한다. 여기에서는 주요 업체를 중심으로 몇 가지 주목할만한 제품이나 기술을 언급하도록 하겠다.

현 시점에서 빅 데이터에 가장 많은 투자를 하고 업체는 IBM이다. 2011년 IBM의 빅 데이터 관련 매출은 11억 달러로서 전세계 IT기업 중 1위를 차지하기도 했다. IBM은 2009년 2월에 스마트 플래닛 전략을 발표하면서 일찍이 BAO(Business Analytics and Optimization)을 강조하기 시작했다. 2010년 7월에 통계 분석 솔루션 업체 SPSS를 12억 달러에 인수했고, 2010년 9월에는 데이터웨어하우스 업체 네티자(Netezza)를 17억 달러

에 인수하기도 했다.

이후 IBM은 2011년 11월에 10PB의 데이터를 수분 내에 분석할 수 있는 'IBM 네티자 하이 캐퍼시티어플라이언스' 발표했고, 2012년 3월에는 의사결정을 위한 예측 분석 역량을 제공하는 'IBM 스마트애널리틱스'를 발표했다. IBM은 지난 5년 동안 140억 달러를 투자해 빅 데이터 관련 업체 24개를 인수하고 8천 여명의 BAO 컨설턴트를 확보해 빅 데이터와 관련된 토털 서비스 제공에 나서고 있다.

EMC의 행보도 주목할 만 하다. 2009년 11월 EMC는 VM웨어, 인텔, 시스코와 함께 VCE(Virtual Computing Environment)를 설립하고 상호 협력하여 사업을 전개하기 시작했다. 그리고 2010년 7월에 데이터웨어하우스 업체 그린플럼(Greenplum)을 인수했으며(인수금액은 미공개), 2010년 11월에는 네트워크스토리지 업체인 아이실론(ISILON)을 22억 5천만 달러에 인수한 바 있다.

EMC는 2012년에 하둡분산파일시스템과 통합된 엔터프라이즈 NAS인 'EMC 아이실론 스케일아웃 NAS'를 출시했다. EMC는 지난 8년간 인수합병에 140억 달러를 투자한 것으로 알려지고 있다. EMC의 3대 슬로건은 Cloud Transforms IT, Big Data Transforms Business, Trust in Your Cloud인데 클라우드와 빅 데이터를 전면에 내세우고 있음을 알 수 있다. 실제로 클라우드와 빅 데이터는 밀접한 연관성을 갖고 있는데 많은 세부 기술들이 겹친다. 클라우드와 빅 데이터는 앞으로 서로 영향을 미치며 동반 발전할 것으로 보인다.

IBM, EMC와 사업 방향은 다르지만 구글은 빅 데이터에서 가장 중요한 업체 중 하나다. 구글은 플랫폼 업체로서 오래 전부터 온라인, 오프라인 할 것 없이 수많은 데이터를 모으고 있는 중이다. 구글은 일찍이 검색을 위해 인터넷 상의 웹 페이지를 수집하고 있을 뿐만 아니라, 지메일, 캘린더 등의 무료 서비스를 통해 사용자의 데이터를 모으고 있으며, 스트리트뷰, 북스 라이브러리프로젝트 등을 통해 오프라인의 데이터를 모으고, 구글 플러스 등을 통해 SNS의 데이터를 모으고, 안드로이드 기기를 통해 디바이스의 데이터까지 모으고 있다.

즉, 사용자가 구글이 제공하는 서비스를 이용하기만 하면 구글의 클라우드에 데이터가

자동으로 쌓이는 구조다. 그리고 구글은 그렇게 모은 데이터를 각종 광고 사업에 활용해 수익을 창출하고 있다. 현재 가장 많은 데이터를 수집하고 있을 뿐만 아니라 가장 다양한 형태의 데이터를 수집하고 있는 업체가 바로 구글이다. 또한 구글은 빅 데이터 관련 기술 및 도구들을 직접 개발하여 사용하고 있는데 여기에서 몇 가지 주요 기술들을 소개하면 다음과 같다.

■ Pregel

- 그래프 알고리즘의 처리를 지원하기 위한 기술로서 1조 개의 데이터를 수 초 내에 처리할 수 있다.
- 참고: <http://googleresearch.blogspot.com/2009/06/large-scale-graph-computing-at-google.html>

■ Dremel

- 대용량 데이터를 분산 처리로 빠르게 분석할 수 있는 기술로서 2006년 이후 구글에서 널리 이용되고 있다.
- 참고: <http://research.google.com/pubs/pub36632.html>

■ Percolator (Caffeine)

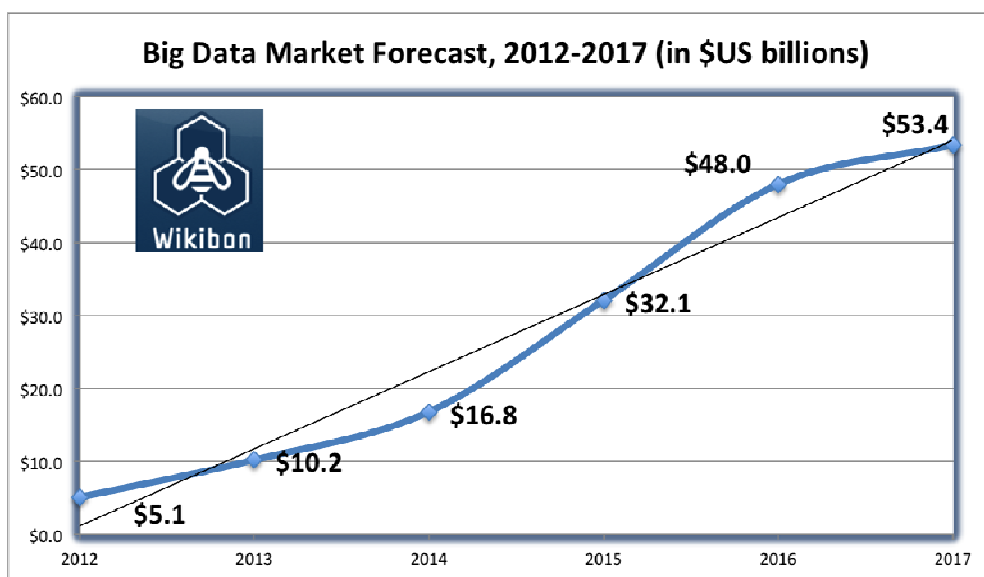
- 검색 인덱스를 작성하기 위한 기술로서 기존의 방법보다 훨씬 신속하게 작업을 처리할 수 있다.
- 참고: <http://research.google.com/pubs/pub36726.html>

세계 최대의 SNS인 페이스북은 그 자체로 클라우드이자 빅 데이터 플랫폼이라고 할 수 있다. 페이스북은 개인의 신상정보 및 관심사, 활동 내역에 대한 각종 데이터를 인터넷에 서뿐만 아니라 오프라인을 통해서도 끝없이 수집하고 있으며 이를 소셜 광고에 활용해 수익을 창출하고 있다. 페이스북의 기업 가치는 바로 이러한 빅 데이터로부터 나오는 것이다.

흥미로운 점은 페이스북이 내부 조직의 프로세스 분석에도 빅 데이터 기술을 적극 활용하고 있다는 점이다. 페이스북은 페이스북에 자사 임직원들이 올리는 글과 타임라인 등을 분석해 서로 커뮤니케이션이 활발한 직원들끼리 팀을 구성하게 하는 등 조직 향상에도 빅 데이터를 활용하고 있다.

3. 빅 데이터 관련 이슈와 전망

빅 데이터를 통해 경쟁에서 앞서 나가려는 기업들의 욕구, 그리고 클라우드의 영향으로 인해 IT 투자가 감소할 것으로 예상되는 상황에서 새로운 시장을 발굴하려는 IT서비스 기업들의 욕구가 빅 데이터의 앞날에 긍정적인 영향을 미치고 있다. 그에 따라 빅 데이터 시장 규모는 향후 5년간 지속적으로 확대될 예정이며, 2017년경에는 530억 달러를 돌파할 것으로 전망되고 있다.



[그림] 빅 데이터 시장 규모 전망 (출처: 위키본)

하지만 빅 데이터는 하드웨어, 소프트웨어, 서비스가 모두 절묘하게 융합돼야만 성과를 낼 수 있을 정도로 고단위도의 역량을 필요로 한다. 그런 이유로 시장조사업체 가트너는 2015년까지 포춘 500대기업의 85% 이상이 빅 데이터를 경쟁우위 확보에 활용하는데 실패할 것으로 예측하기도 했다. 이에 대해 부연하자면, 빅 데이터를 도입하는 기업이 많을 지라도 이를 제대로 활용해 유의미한 성과를 내는 기업은 적을 것이라는 뜻이다.

그 이유는 무엇일까? 빅 데이터를 제대로 활용해 성과를 내기 위해서는 그만큼 성숙된 IT/조직 문화를 갖추고 있어야 한다. 정보시스템조차 겨우 운용하고 있는 기업이 빅 데이터로 성과를 낼 리 만무한 것이다. 또한 IT와 비즈니스 도메인에 대한 지식과 경험을 갖춘 인력을 확보하는 것이 필요하다. 즉 기술 활용 능력과 더불어 제조, 소매, 의료 등 특

정 업종의 사업 메커니즘과 비즈니스 룰, 업무 프로세스 등 도메인 지식(Domain Knowledge)을 갖춘 인력을 갖추어야 한다. 그래야 유의미한 비즈니스 인사이트를 도출할 수 있기 때문이다. 또한 통계학 및 수학적 지식을 갖춘 데이터 사이언티스트(Data Scientist)의 확보가 필수적이다. 그러한 이유로 IBM은 빅 데이터 시장에서 서비스 역량의 확보가 중요하다고 보고 컨설팅이 가능한 전문 인력을 집중적으로 양성하고 있는 것이다.

또한 추가적으로 다음과 같은 이슈들도 존재하는데 이에 대해서는 앞으로 지속적인 논의가 필요할 것으로 보인다.

- 개인정보 유출 및 사생활 침해 문제: 생데이터를 기반으로 하기에 개인의 사적인 데이터가 그대로 들어있으며 이를 적절히 필터링하고 보호해야 한다.
- 보안 및 영업비밀의 유출 문제: 데이터 분석 및 활용에 외부 업체의 도움을 받을 시 문제가 될 수 있다.
- 데이터의 오용 및 부적절한 이용 문제: 이 부분을 간과하면 고객에게 부적절한 상품을 추천하거나 불쾌감을 주는 등의 문제가 발생할 수 있다.

지금까지 살펴본 내용을 바탕으로 향후 빅 데이터 시장의 전개방향을 정리해보면 다음과 같다.

첫째, 데이터의 가치가 증대됨에 따라 업체들간에 데이터를 거래하는 데이터 마켓플레이스가 주목 받게 될 것이다.

어떤 데이터를 낮은 비용으로 획득할 수 있는 사업자와 그 데이터를 효과적으로 이용할 수 있는 사업자가 반드시 일치하지는 않는다. 시장에는 이미 데이터의 판매자와 구매자가 만나 거래를 할 수 있는 데이터마켓플레이스가 등장하기 시작했다. 예를 들어, 인포침스(Infochimps)는 GIS, SNS 등의 각종 데이터를 상용으로 판매하고 있는 중이다.

둘째, 물리적 현상을 나타내는 데이터를 수집함에 따라 온라인과 오프라인의 연계가 더욱 중요해질 것이다.

빅 데이터는 인간의 조작을 통해 발생한 데이터뿐만 아니라 M2M(Machine-to-Machine)을 통해 각종 센서나 디바이스로부터 데이터를 획득하는 것도 포함한다. 디바이스와 그 디바이스가 취급하는 데이터 사이에는 높은 연관성이 있기 때문에 데이터를 수집하는 소스는 앞으로 계속 늘어나게 될 것이다. 빅 데이터에서는 다양한 유형의 데이터를 수집하는 게 중요한 경쟁력이기 때문이다.

셋째, 빅 데이터에서 탁월한 성과를 내기 위해서는 무엇보다도 빅 데이터를 활용할 수 있는 성숙된 IT/조직 문화의 확립해야 한다. 더불어 IT 및 비즈니스 도메인 지식을 갖춘 인력, 그리고 통계학, 수학적 지식을 갖춘 데이터 사이언티스트를 확보하는 것이 중요한 선결과제라고 볼 수 있다.

만일 그러한 환경이 제대로 조성되지 않은 채로 어설픈게 최신 빅 데이터 기술과 도구를 도입해봤자 비즈니스에는 별반 도움이 되지 못한 채 투자 비용만 날릴 가능성이 크다. 국내 환경에서 빅 데이터 트렌드와 관련해 가장 우려되는 점이 바로 이 부분이다.

빅 데이터는 예술로 치자면 종합예술에 가깝다. 수많은 개별 기술로 구성돼 있을 뿐만 아니라 각각의 기술적 깊이도 상당하며, 기존의 엔터프라이즈 기술을 훨씬 능가하는 대용량 처리를 다루기에 많은 투자를 필요로 한다. 또한 클라우드, M2M과도 밀접한 관계를 맺고 있다. 거기에다 기술 역량 이상으로, 비즈니스 목표를 수립하고 비즈니스 인사이트를 찾아내고 비즈니스에 접목하는 서비스 역량이 중요하다.

한마디로, 빅 데이터로 성과를 내려는 기업은 먼저 그럴만한 자격을 갖추어야 하는 것이다. 단지 기술 도입만으로 성과를 낼 수 있다는 생각은 버려야 하며, 조급하게 생각해서도 안 된다. 빅 데이터를 제대로 활용할 수 있는 인프라를 갖추고 전문 인력을 확보한 기업만이 빅 데이터를 통해 만족스러운 결과를 얻을 수 있을 것이다. @