데이터 관점에서 바라보는 빅데이터와 공개 SW

정휘웅* 박준석* 박성호** 김태열***

본 고에서는 데이터 관점에서 바라보는 빅데이터 처리, 그리고 이를 위한 공개 SW의 역할에 대해서 기술한다. 지금까지 시스템 관점 및 인프라 스트럭처 관점, 처리 절차 관점에서 빅데이터를 이야기 했다면, 이제는 데이터 자체의 처리 및 분석으로 주제가 세분화되어 가고 있다. 그러나 세부적인 데이터 처리를 위해서 어떤 형태의 소프트웨어를 어떻게 구성할 것이며 어떻게 환경을 마련해야 할 것인가에 대해서는 제대로 제시되지 않고 있다. 본 고에서는 이를 해결하기 위한 효과적인 방법이 공개 SW를 도입하는 것이며, 이를 위해 어떤 환경이 갖추어져야 하는가에 대해서 기술하고자 한다.

목 차

* NIPA 공개 SW 역량프라자/선임연구원
** NIPA 공개 SW/수석연구원
*** NIPA 공개 SW 지역팀/팀장

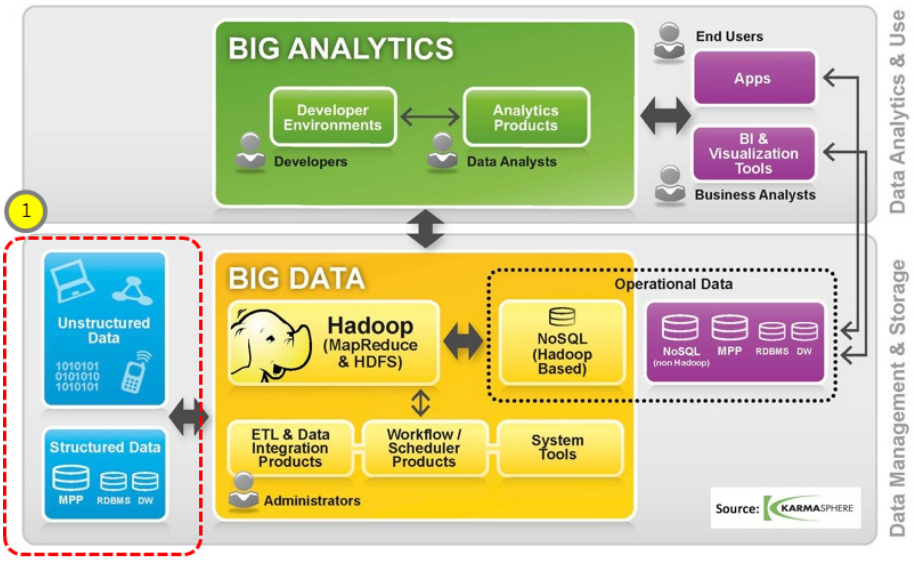
Ⅰ. 서 론

최근 빅데이터와 관련된 이슈는 플랫폼과 처리 절차를 넘어서 분석에 대한 언급이 주류를 이루고 있다. 그러나 업무 담당자들은 빅데이터로 볼 수 있는 정보의 범위에서 처리하는 방법, 도입 절차에 이르기까지 실무적인 부분에 있어서 접근 방법을 잡기 어렵다. 요구사항이 추상적이며 내부적으로 보유한 데이터를 바탕으로 원하는 결과를 도출하기에는 정보가 부족하거나 구조화가 덜 되어 있는 경우가 많기 때문이다.

게다가 각 레거시 시스템들에 저장된 데이터들은 다양하고 산업군별로 수집되는

형태의 데이터도 다양한데도 불구하고, 빅데이터에서 이야기하는 데이터는 이 모두를 하나로 묶어서 이 자료에서 무엇인가 의미 있는 것을 추출할 수 있다는 추상적인 희망을 가지고 있다. 그러나 빅데이터의 분석을 제대로 수행하기 위해서는 실제 어려움이 많이 따른다. 이를 시스템의 관점이 아닌 데이터의 관점에서 바라보게 되면 빅데이터를 다루는 것이 정보의 개별 형태에 따라 다른 특징을 가지고 있다는 것을 발견할 수 있다. 가령 미항공우주국에서 이미 활용하고 있으며 천문 관측에 있어 지구로 근접하는 NEO(Near Earth Object)를 추적하는 빅데이터 분석 시스템[1]을 개발한다고 생각해보자. 매일 고성능 관측 카메라가 촬영하는 막대한 용량의 이미지 사이에서 발생하는 미세한 밝기 차이를 추적하여 그 때에 발생하는 정보를 분석할 수 있어야 한다. 이 경우 사진 정보 혹은 로그 정보, 좌표 정보들은 플룸(Flume), 스쿱(sqoop), 척와(chukwa)와 같은 빅데이터 수집기, 하둡(Hadoop)과 같은 기반 파일 시스템 환경, 하이브(Hive)와 같은 질의(query) 환경이 설치되었다 하더라도 사진 사이의 차이점을 찾아내는 알고리즘은 어떠한 상용 패키지나 공공 SW 프로젝트에서도 지원하지 않고 있다. 맵리듀스(MapReduce) 과정을 통해 데이터를 처리하기 위해서도 텍스트 이외의 정보들은 각각의 형태에 적합한 정보가 제공되어야 빠른 처리가 가능하다.

이러한 문제를 해결하기 위해 미국 오바마 행정부는 빅데이터와 관련된 다양한 툴킷과 관련된 기술들을 공개소프트웨어 기반으로 추진하고 있으며[2], 빅데이터의 수집과 분석



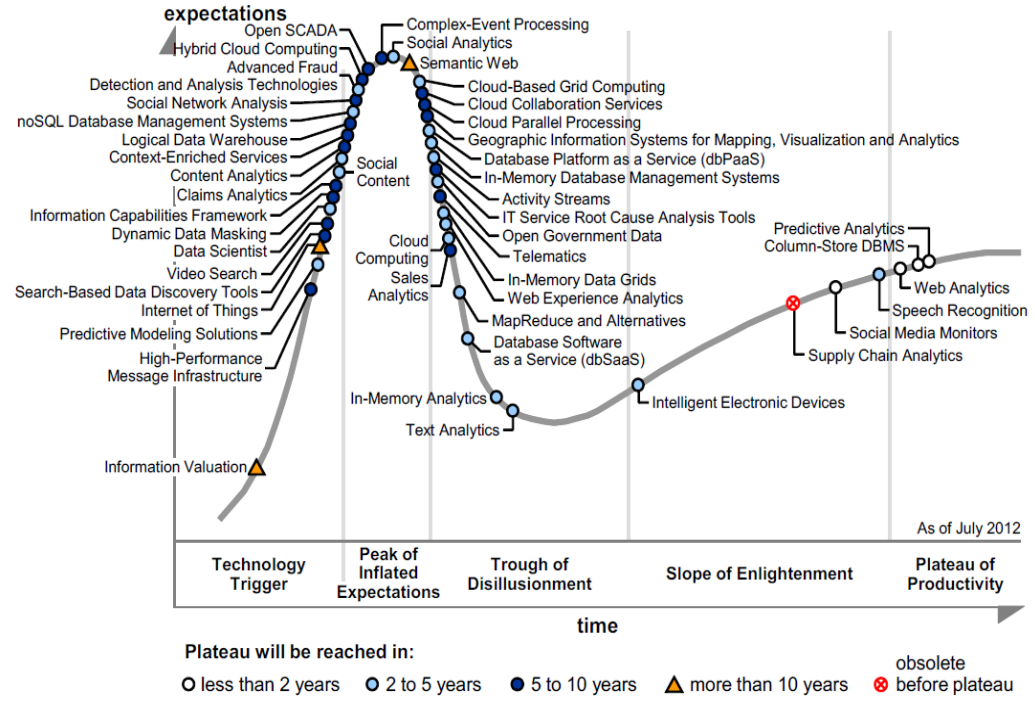
(그림 1) 빅데이터 처리 구조[4]

은 향후 중요한 이슈로 제기될 것이다. 빅데이터의 이슈가 이처럼 세부적으로 분화되기 시작하면서 2012 년 가트너 그룹은 빅데이터에 대한 Hype Cycle 을 발표하였다[3]. 이 보고서는 지금까지 솔루션 및 인프라 구조로 빅데이터를 바라보는 것을 보다 세분화하여 접근하고, 이에 대한 개별적 분류 작업을 한 것에 대해 의의를 둘 수 있다. 본 고는 이 중 데이터 분석에 대한 관점보다는 빅데이터에서 어떤 정보를 어떻게 추출할 것인가를 중점적으로 이야기 할 것이다.

2 장에서는 데이터 형태의 관점에서 빅데이터를 분류하고, 3 장에서는 각각의 분류된 빅데이터가 실제 수집되고 어떻게 처리되어야 하는지에 대한 공개 SW 의 관점과 4 장에서는 앞으로 나아가야 할 방향과 정책적 필요성을 제시하였다.

II. 데이터 관점에서 바라보는 빅데이터

가트너 그룹이 발표한 2012 년 7 월 기준 빅데이터 Hype Cycle 에는 빅데이터와 관련된



Source: Gartner (July 2012)

(그림 2) 가트너 그룹이 발표한 빅데이터 Hype Cycle(2012 년 7 월)[3]

여러 가지 이슈들이 제시되고 있으며, 이중 주목해야 하는 부문은 정보 가치화(Information Valuation)이다[3]. 정보 가치화 부문은 기존 자료들이 정확하고 정교한 정보를 담음으로써 실제 빅데이터 분석 시 틀리지 않은 정확한 정보를 관계성에 기반하여 제공할 수 있도록 하는 것을 목적으로 하고 있으며, 향후 10 년 정도의 성숙 기간을 예견하고 있다. 데이터가 정확하지 않으면 빅데이터의 분석이 어려우며, 국내에서는 데이터베이스진흥원을 중심으로 데이터의 품질 관리와 절차의 표준화를 위한 ISO-8000 기반 표준화도 이루어지고 있으나 빅데이터 기반 자료로서 데이터 가공과 관리는 그만큼 시급한 주제로 대두되고 있다[5].

빅데이터의 출발점은 데이터의 처리이며, 제대로 된 데이터를 추출할 수 있어야 제대로 된 빅데이터 분석을 할 수 있다. 몇 가지 데이터 유형을 예로 들어 처리 방법과 추출할 수 있는 정보, 응용 방법에 대해 살펴보도록 하겠다.

1. 지리 기반 정보

지리 기반 정보는 GPS 에서 제공하는 정보도 있으나 해당 위치의 행정구역에서 주는 정보가 중요하다. 하나의 측정된 데이터는 여러 가지 메타 정보를 제공할 수 있다. GPS 정보 하나만으로는 세부적인 정보를 제공할 수 없으며, 여기에 연관된 키워드 정보를 함께 제공할 수 있어야 한다. 지리 기반 정보로 제공하는 서비스(Location Based Service: LBS)는 시간 정보와도 밀접한 연관성을 띤다. 가령 전자 발찌를 찬 성폭력 범죄자의 이동 경로를 추적하기 위해서는 그 때 그 때 위치를 탐색하는 것이 아니라, 특정 시간대별로 이동하는 경로를 추적할 수 있어야 한다. 아울러 이 정보들과 신상 정보, 해외의 유사한 사례 등을 종합적으로 분석해야 한다[6].

2. 수치 및 센서 정보

수치 및 센서 정보(numeric, sensing data)는 빅데이터에서 분석하기가 매우 어려운 자료다. 특히 미래 정보에 있어서 막대한 분량으로 제공될 정보가 바로 수치 및 센서 정보다. 수치 정보는 증권 거래에서 오는 지수, 종가 정보, 환율 교환 정보, 국제 시장에서 발생하는 다양한 원자재 가격 동향 정보, 각 국가별 기준금리에서부터 SCADA(Supervisory Control And Data Acquisition) 시스템에서 수집하는 각종 생산 공정 데이터 등 다양한

형태의 자료들이 이에 해당한다[7]. 수치 및 센서 정보는 다양한 데이터 마이닝 기법을 통해 가공될 수 있으며, 단순히 축적되는 형태가 아닌 각 수치 정보가 의미하는 바를 예측하고 추론하는 과정이 선결되어야 한다[8]. 여기에는 해당 정보가 실제로 유의미한 정보인지 아닌지를 감별하는 알고리즘이 함께 구축되어야 한다[9].

3. 소리 정보

소리 정보(Sound, Voice, Noise)는 우리가 생각하는 음성인식의 관점으로 볼 수 있으나 소리는 그 이상의 범주로 보아야 한다. 사람의 입에서 나오는 소리 역시 언어 정보도 있으나 웃음, 휘파람, 울음과 같은 형태의 정보들이 있으며, 기계의 경우 특정한 문제점을 표현하면서 사람이 들을 수 없는 저주파 혹은 고주파 정보들이 존재한다. 빅데이터는 이러한 정보를 지속적으로 수집하여 파형 정보로 저장하고, 각 상태의 정보를 패턴화하여 분석할 수 있어야 한다. 음악 정보의 경우 특정 부분의 음악 톤(tone)을 수치화하여 하나의 패턴 정보로 저장하고 음악 검색 환경에 활용할 수 있다. 가령 서치 바이 허밍(Search by humming)과 같은 검색 환경에 소리 정보의 대규모 분석이 필요할 것이다[10].

4. 텍스트 정보

텍스트 정보는 소셜 네트워크에서 실시간 벌어지고 있는 현상과 지금까지 언급된 형태의 정보에서 미래 추세를 예측하는 것에 이르기까지 다양한 부문으로 응용이 이루어져야 한다. 텍스트 정보는 지금까지 가장 많이 분석이 되어 왔고 빅데이터에 있어서도 가장 핵심 자료로 분류되고 있음에도 불구하고 가장 분석이 적은 분야이다. 빅데이터에서 핵심으로 삼는 양(volume), 속도(velocity), 다양성(variety) 측면에 있어서 텍스트 정보는 가장 다양한 형태의 특성을 띤다[11].

양과 속도에 있어서는 트위터, 페이스북과 같은 소셜 네트워크 정보가 가장 큰 범주를 차지하고 있다. 그러나 양과 속도 측면에 있어서 상대적으로 속도가 느리다 하더라도 신문기사, 논문, 특허, 법령 정보, 기업 내부의 문서 등 다양한 형태의 정보들이 빠른 속도로 증가하고 있다. 이러한 정보들은 양과 다양성, 복잡도는 소셜 네트워크에 비해 보다 높은 수준의 언어처리기술을 요구한다. 맥락기반 정보 및 의미기반 정보, 온톨로지 정보 등 다양한 형태의 의미 정보들이 텍스트 정보를 형성한다[12].

5. 이미지 정보

모바일 기기가 보편화 되고 지리정보와 연동되는 이미지 정보가 급속도로 증가함에 따라 이미지 정보를 처리하는 기술들 역시 시장에 다양하게 소개되고 있다. 특히 이미지 정보 내에서 특정 사물 정보를 찾아내거나 인물의 안면을 인식하는 핑거프린팅(fingerprinting) 기술은 앞으로 이미지 정보가 증가할수록 그 양이 급속히 늘어날 수 있다[13]. 이미지 정보는 기상, 의료, 군사 분야 등에서도 빅데이터를 이용하여 직접적인 처리를 할 수 있다. 이미지 사이의 차이점, 이미지 내에 존재하는 사물의 모양을 인식하는 등 다양한 형태의 정보를 추출할 수 있다.

6. 동영상 정보

동영상 정보는 음성과 장면, 캡션과 같은 다양한 정보들이 연동되어 추출될 수 있다. 동영상은 각 단계별로 주요 장면을 캡처하여 이미지 정보 처리 영역으로 할 수 있을 뿐만 아니라, 뉴스 캡션과 뉴스 동영상 사이의 음성 정보를 동기화하여 음성인식시스템의 성능을 향상시키는 자료로 활용할 수도 있다. 아울러 동영상에서 제공하는 다양한 형태의 주요 배경 화면을 바탕으로 사용자가 촬영한 주변 풍경과 관련 있는 정보들을 효율적으로 인덱싱하고 검색할 수도 있다[14].

7. 온톨로지 및 의미, 관계 기반 정보

앞서 제시한 모든 형태의 정보들은 하나의 메타 정보 형태로 저장되나, 독립된 형태로 저장되면 빅데이터에서 제대로 된 의미 정보를 추출할 수 없다. 모든 형태의 데이터를 가지고 있다 하더라도 각 정보들 사이의 연관성이 없으면 빅데이터 가공을 위해서 정보를 사용할 수 없다. 가령 이미지 정보를 수집하는 경우에도 해당 이미지는 촬영한 날짜, 장소, 그리고 메타 정보로써 이미지 내에 촬영된 물체의 이름, 태그된 인물의 정보에 이르기까지 다양한 의미 정보들이 포함될 수 있다. 각 정보들은 다시 상위계층 온톨로지와 연계되어 해당 사진이 촬영된 곳의 정보를 알려주거나 이와 관련된 음악 정보를 연관 지을 수도 있으며, 역사적인 사실 정보와 연계할 수도 있다. 이러한 정보들은 대개 트리플(triple)이라는 RDF(Resource Description Framework) 기반의 정보로 저장되는데, 빅데이터 부문에 있어서도 이 정보는 소셜 네트워크 정보가 늘어나고 분석 데이터가 늘어날수록 그 복

잡도가 늘어난다[15]. 따라서 다양한 사용자 요구사항을 반영하고 처리하기 위해서 온톨로지 기반 정보는 빅데이터에서 해결해야 하는 부분이다.

III. 빅데이터 가공과 공개 SW

1. 데이터 가공을 위한 기술

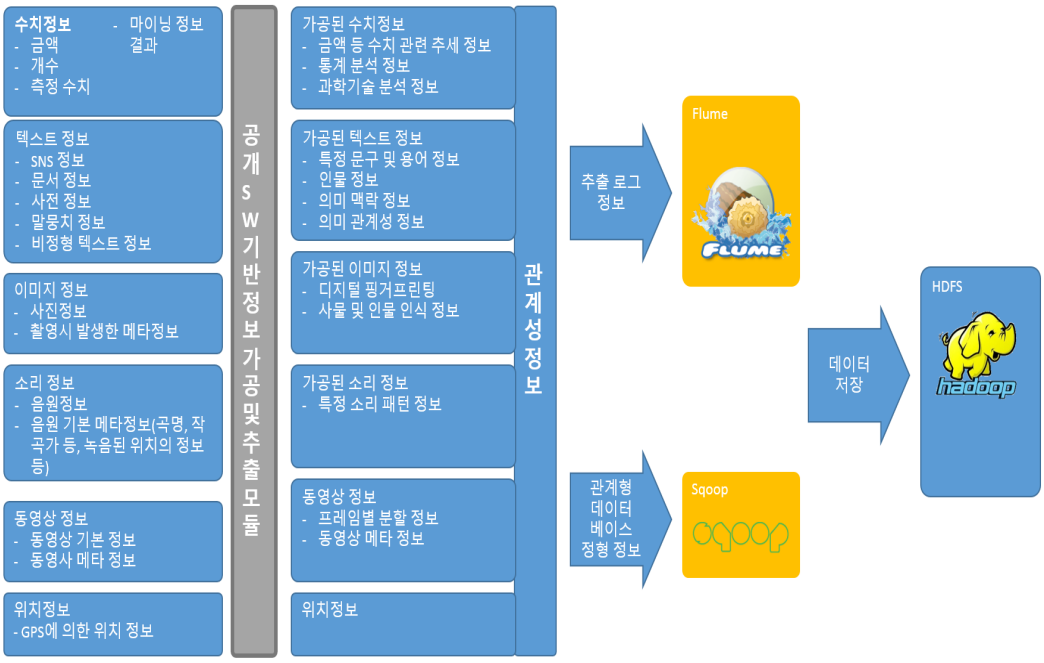
오바마 정부는 XDATA 프로그램을 공개 SW 기반으로 툴킷을 제공할 것으로 알려졌다. 이는 빅데이터의 특성이 유연하고 가공을 위해서는 다양한 형태의 변환기들이 요구되기 때문이다. XDATA 프로그램은 4 년 동안 연간 2,500 만 달러를 투자해 대용량 데이터를 분석할 수 있는 컴퓨팅 기술과 소프트웨어를 개발하는 것으로 알려져 있으며, 분석 대상 데이터는 준(semi)정형 데이터(표, 관계형, 카테고리형, 메타데이터)와 비정형 데이터(텍스트 문서, 메시지 전송)를 모두 포함하고 있다[16].

이처럼 동일한 형태의 정보라 하더라도 그 특성에 따라서 시스템은 다양한 형태의 정보 추출, 축적, 분석 작업을 수행해야 한다. 각 정보들 사이에는 관계성을 가지고 있어야 하며, 각 정보들은 사용자에게 의미 있는 정보로 제공되기 위한 기초 자료를 제공할 수 있어야 한다. 가령 지리 정보와 연계된 이미지 정보, 소리 정보 추적 등의 시스템은 야생 동식물의 이동 경로 혹은 서식지를 확인하는데 응용할 수 있으며, 기가픽셀 카메라와 같은 초고화질 카메라 정보를 이용한 범죄 예방 시스템은 실시간으로 촬영되는 광역 정보의 데이터를 실시간으로 분석하여 범죄 예방 혹은 화재, 산불 감시 등의 환경에 응용할 수 있을 것이다.

그러나 빅데이터 분석을 위해서는 우선 시스템이 충분한 추론을 할 수 있는 학습 데이터가 있어야 하며, 이 데이터에서 추출할 수 있는 의미 있는 데이터를 확보한다는 것을 전제로 한다. 이 경우에는 내부 정보를 수집하여 하둡 파일 시스템(HDFS)에 저장할 수는 있으나 수집된 자료로부터 필요한 정보를 추출하는 것은 개별적인 기술들이 적용되어야 한다.

2. 정보 가공을 위한 공개 SW 기반 프로젝트 제안

추출해야 하는 데이터의 종류가 다양하고 변동성이 다양한 경우에는 상용 솔루션을 구축하는 데에는 많은 시간이 소요될 뿐만 아니라, 상업적으로 성공을 보장할 수 없기 때문



(그림 3) 공개 SW 기반으로 구성되는 빅데이터 처리 및 수집

에 개별적인 모듈을 기업들이 구축하는 것을 기대하기 어렵다. 이러한 경우 공개 SW 기반 정보가공 및 추출 모듈을 개발하고, 각 정보에서 추출된 데이터를 바탕으로 플룸이나 스쿱과 같은 수집 시스템을 통해 하둡 파일 시스템(HDFS)으로 자료를 저장할 수 있다.

공개 SW 는 비용이 절감되고 기술적 종속성이 없어서 많은 국가에서 공개 SW 를 국가적인 프로젝트로 지정하여 추진하고 있으며, 특히 공개 SW 는 다음과 같은 이점을 가지고 있다. 첫째, 소스코드에 대한 자유로운 접근으로 최신 기술을 쉽게 습득할 수 있다. 운영체제의 핵심 기술, 대용량 분산 처리의 핵심 기술, 정보 인덱싱 등 여러 전문가들의 기술을 습득할 수 있다. 둘째, 창의적 아이디어 수용 등 기술혁신 이점이 있다. 기업 관점에서 혁신적 신기술이라 하더라도 비용적인 측면을 고려하여 실제 솔루션에 도입되는 시기가 늦어지는데 반해 공개 SW 는 그 때 그 때 신기술 아이디어를 도입할 수 있다. 셋째, 개발 기간이 단축되고 비용이 절감된다. 여러 사용자들이 소스코드에 대해 다양한 형태의 튜닝 및 개발 과정을 거쳐 테스트 및 인터페이스 설계 등에 들어가는 비용을 줄일 수 있다. 넷째, 시장 경쟁 촉진 및 다양성을 확보할 수 있다. 상용 SW 가 주도하는 시장에 있어서 공개 SW 의 등장은 좋은 자극이 되며, 상용 SW 의 기술 개발과 시장 확대 등을 촉진할

수 있으며, 소비자들에게는 다양한 제품 선택의 기회를 제공한다[17].

이처럼 다양한 이점이 있기 때문에 다양한 형태의 정보를 분석하고 추출하며 처리하기 위해서는 공개 SW 기반 솔루션 접근이 매우 적절하다. 상용 SW 개발 기업들이 개별적인 컴포넌트에 상업적 목적으로 투자하기는 매우 어렵다. 공개 SW 를 개발하는 다양한 커뮤니티에 의해서 데이터의 특성에 맞는 프로젝트를 만들고, 이를 통해 다양한 형태의 정보 분석이 이루어지도록 정책적인 지원이 이루어지는 것이 보다 타당할 것이다. 미국의 경우에도 앞서 언급한 바와 같이 XDATA 프로젝트를 공개 SW 프로젝트 형태로 제공하고, 이를 통해 다양한 형태의 빅데이터 자료들을 분석하는 계획을 수립하고 있다.

IV. 결 론

빅데이터의 분석이 효율적이고 정확한 통찰력을 제공하기 위해서는 빅데이터에서 의미 있는 정보를 추출하는 과정도 분석만큼 중요하다. 오바마 선거 캠프의 데이터 분석 전략이 성공했던 이유 역시 2008 년 선거 자료를 충실히 수집하고 각 정보 사이의 연관성을 적절히 지워 두었으며, 유권자들의 성향을 세분화 하였기 때문에 성공적인 분석을 수행할 수 있었다. 이에 반해 아직까지 빅데이터로 볼 수 있는 자료들은 이처럼 특정 목적에 맞도록 가공되어 있지 않으며, 상호 연관성도 낮아서 실제 목적에 맞도록 응용하기 위해서는 앞으로도 수행해야 할 길이 매우 많다. 따라서 출발점으로 볼 수 있는 데이터의 가공, 추출, 연관성 부문에 있어서 많은 투자가 있어야 하나 기업이나 공공부문 한 곳에서 모든 것을 전담할 수 없다. 따라서 미래 빅데이터 전략에 있어 모든 목적에 맞는 다양한 형태의 데이터 처리 엔진을 구축하기 위해서는 공개 SW 와 같은 개방형 구조가 효율적인 방법으로 그 역할을 할 수 있을 것이다.

아울러 공개 SW 가 최근 추세인 공개 데이터 부문과 연계된다면 빅데이터 정보의 수집 및 처리가 더욱 빨라질 것이며, 이를 통한 분석도 더욱 높은 수준에서 통찰력을 제공할 수 있을 것이다. 기계에 의해 처리되는 고급 정보를 바탕으로 더욱 많은 정보 분석 인력에 대한 수요가 일어날 수 있으며, 자연스럽게 일자리 창출과 같은 다양한 부문의 선순환적인 파급 효과를 가져오게 될 것이다.

<참 고 문 헌>

[1] NASA, Near Earth Object Program, http://neo.jpl.nasa.gov/

[2] US White House Immediate Release (2012) "OBAMA ADMINISTRATION UNVEILS "BIG DATA" NITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS", 2012.

[3] Lapkin, Anne, "Hype Cycle for Big data, 2012," Gartner, July, 2012.

[4] Karma Sphere, http://www.karmasphere.com

[5] ISO/TS 8000-150:2011 Data quality - Part 150: Master data: Quality management framework - http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=54579

[6] 정구민, 최원석, "스마트폰 위치기반 서비스(LBS) 기술동향", TTA Journal No.130, 2011, pp.75-81.

[7] SCADA, http://en.wikipedia.org/wiki/SCADA

[8] "Enter the Big Data Matrix: analyzing meanings and relations of everything", MSDN, 2013, http://blogs.msdn.com/b/hpctrekker/archive/2013/04/07/enter-the-big-data-matrix-analyzing-meanings-and-relations-of-everything-1-2.aspx

[9] Christian Bizer, "The Meaningful Use of Big Data: Four Perspectives–Four Challenges", SIGMOD Record, Vol.40, No.4, December 2011, pp.56-60.

[10] Nauman Ali Khan, "Hybrid Query by Humming and Metadata Search System(HQMS) Analysis over Diverse Features," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol.2, No.9, 2011, pp.58-66.

[11] Big data spans four dimensions: Volume, Velocity, Variety, and Veracity, IBM, http://www-01.ibm.com/software/data/bigdata/, 2012.

[12] Paul Viola, "Learning to Extract Information from Semi-structured Text using a Discriminative Context Free Grammar", SIGIR 2005 Draft.

[13] 최재귀, 이경현, "Design of Digital Fingerprinting Scheme for Multi-purchase", 한국멀티미디어학회지, 7(12), 2004, pp.1708-1718.

[14] "웹상의 동영상 콘텐츠들에서 사용자 지정 부분만을 복사 없이 조합, 새로운 동영상을 생성재생 서비스하는 소프트웨어 개발 과제", 특허청 R&D 특허센터, 특허기술동향조사 보고서, 2011.

[15] Jans Aasman, "Triple stores or nosql in the enterprise", Franz, 2008, http://franzdownload.com/

[16] "미 정부, 빅데이터에 2 억달러 투자…오바마 "모두 도우라"", 전자신문 2012, http://www.etnews.com/news/international/2575374_1496.html

[17] "2012 공개소프트웨어백서", 정보통신산업진흥원, 2013, pp.3-8.

* 본 내용은 필자의 주관적인 의견이며 NIPA 의 공식적인 입장이 아님을 밝힙니다.