한국컴퓨터정보학회지
제20권 제2호, 2012. 12.

Big Data의 이해 - 가치와 도입전략

• 이성희(LG엔시스 CTO)

I. 서론

하루에도 수많은 Data가 만들어지고 있다. 60초안에 1억 6천 8백만개의 E-mail이 발송되고, 51만개의 facebook comment가 달리며, 1만 3천개의 iPhone 어플이 다운로드되는 등 소셜네트웍을 통한 Data가 기하급수적으로 생성되고 있다. 정보시스템의 고도화, 모바일, Cloud, 소셜네트웍스의 일상화로 인해 생성되는 Data의 양이 제타바이트시대에 돌입하고 있다. 더불어 정형 Data보다는 비정형Data 폭증으로 전세계 데이터량은 매년 40%씩 증가하고 있다.

이렇게 증가하는 Data를 분류해 보면 정형Data가 15%, 비정형Data가 85%로 나누어진다. 정형Data는 기존 Database 관리시스템인 Oracle, DB2, SQL등으로 관리되어 왔고 분석 도구 역시 잘 알려진 Data Mining 분석 Tool을 통하여 활용되어 왔으며 장비 역시 Global Vender의 제품들이 많이 출시되어 있다.

그러나 facebook, twitter, youtube등에서 촉발되는 85%에 이르는 비정형데이타는 그 망이 기하급수적으로 늘 뿐 아니라 여러 가지 형태로 발생하기 때문에 데이터를 분석하는 것은 매우 어렵다.

다시 말하면 기업이나 조직에서 분석에 필요한 데이타가 기존 데이타기술로는 관리할 수 없는 데이타로서 새로운 분석 도구를 필요로 한다. 특히 빅데이타 중에서도 비구조적인 데이터로서 첫째, 로그파일같은 반구조데이타, 둘째, 데이터셋이 하나이상의 구조를 갖는 데이터, 셋째, 하나의 데이터세트가 어떤 구조도 갖지 않는 데이터로 구분할 수 있다.

그리고 사회적인 분류로 보면 오픈데이타와 소셜데이타가 있다. 오픈데이타는 기상정보, 교통정보 등 정부, 공사 또는 지방자치단체 등이 보유하고 있는 공통데이터베이스로서 일반에 공개된 데이터이며, 소셜데이타는 소셜미디어 즉 페이스북이나 트위터등에 개인 또는 법인이 올린 데이터로서 그 분석의 가치가 매우 중요시되고 있다. 또한 빅데이타는 매년 60%정도 증가될 것으로 예상되고 있으며, 축적된 데이터활용 능력이 중요한 과제로 대두되고 있다. 2020년까지 35 zetta byte 까지 폭증할 것으로 예상되고 있으나, 그에 비해 3분의 1정도만이 분석할 수 있는 능력이 없을 것으로 보여진다.

II. 본론

1. 빅데이타의 특징

빅데이타의 특징으로 5가지를 들 수 있는 데 첫째, Volume으로서 기존 DB보다는 규모가 훨씬 크고 일정 기준으로 구분하지 않는다. 둘째, Velocity로서 배치, 리얼타임, 스트림형태, 실시간분석과 반응을 필요로 한다. 셋째, Variety로서 구조적 데이타와 비구조적데이타를 포함한다. 다양한 구조의 데이터를 서로 연관해서 분석할 수 있어야한다. 넷째, Complexity로서 위의 3가지 특성에 따라 보관, 운영 활용하는 것이 매우 복잡하다. 마지막, Value로서 기존 구조적데이타는 거래를 안전하게 처리하기 위한 목적이지만 이는 경쟁력 및 운영효율성에 직접 큰 영향을 줄 수 있다. 따라서 빅데이타는 규모가 크고, 빠르며, 다양해서 복잡하지만 큰 가치를 가지고 있다.

2. Analytics의 중요성

리서치기관의 조사에 따르면 경영성과가 높은 기업이 낮은 기업보다 Analytics를 더 많이 활용하고 있다는 통계가 있으며 활용분야에서도 재무관리에서부터 영업/마케팅/고객관리는 물론이고 인력관리에까지 활용을 넓히고 있다. 특히 운영효율성, 전략수립, 고객서비스에서도 높은 활용도를 보이고 있다.

3. 분석 활용 대상

분석활용은 업종에 따라 다르며, 새로운 분석을 발견하기 위해 노력이 많아지고 있다. 실시간분석으로는 은행의 신용위험 및 시장 위험 분석, 은행의 부정사용 및 자금세탁 탐지, 금융 및 통신회사의 이벤트마케팅, 유통업종의 마크다운 최적화, 공공분야의 보상 및 과제 부정청구등으로 들 수 있다. batch성 분석으로는 항공회사의 예방정비, 소셜미디어 감성분석, 제조업체의 수요예측, 전자의료기록관리의 질병분석, 전통적 데이터웨어 하우징, 마이닝테스트, 비디오감시 분석 등이 있다. 이러한 분석의 예는 다음장에서 살펴보도록 한다.

4. 빅데이타의 가치

빅데이타의 가치는 크게 두가지로 나누어 볼수 있다. 첫번째로 Agility를 들 수 있는데 이벤트감지, 데이터 확보 분석수행 의사결정, 행동착수라는 일련의 행동 과정을 빅데이타 분석을 빠르게 수행하여 경쟁우위 요소를 가질 수 있게 한다. 둘째로 Relevance를 둘 수 있는데 VOC감성분석, 위치정보 연계, 웹로그분석을 통한 고객구매심리파악, 장바구니분석을 통한 구매 포기 요인 파악, 센서 행동 분석 기반으로 고객상황 인지를 통하여 연관성에 기반한 가치있는 제안을 가능케 한다. 따라서 높은 고객만족, 재구매 및 반복구매, 고객이탈방지와 같은 가치있는 행위를 유발하게 할 수 있다.

5. 적용 효과

국내 적용효과는 아직 나타난 통계가 없으므로 글로벌 적용 효과를 살펴보면, 미국의 의료산업에서는 1년간 3천억달러 가치가 발생하며, 년간 생산성 0.7% 향상한다고 한다. 유럽의 공공행정분야에서는 1년간 2천5백억유로가치가 발생하며 년간 생산성이 0.5%향상되며, 글로벌 개인위치 데이터는 년간 1천억달러 이상 수익을 나타내며, 사용자 가치가 7천억달러에 달한다고 한다. 미국 유통업종에서는 순매진이 60% 이상 증가하며, 년간 0.5~1.0% 생산성이 향상되고 제조업종에서는 제품개발/조립원가가 50% 절감되며, 운전자본이 7%가 절가된다는 통계들로 맥킨지에서 보고하고 있다.

6. 기술과 오프소스

앞에서 살펴본 바와 같이 분석의 중심이 되고 있는 데이타가 기존의 것과 다른 특성을 가지고 있으므로 이에 관련된 기술 역시 새로운 것이다. 사실 빅데이타도 기술의 발전에 따라 가능하게 된 측면이 많다. 특히 그 중심에는 Hadoop Ecosystem이 있다. 또한 데이터의 수집, 저장, 분석, 표현에 이르기까지 일련의 process에서 사용되어지는 오픈소스의 솔루션에 대한 이해와 적용 기술이 매우 복잡하다.

첫째, 데이터 수집은 데이터 발생원으로부터 안정적인 저장소로 저장하는 기능을 수행하는 것으로 대표적인 오픈소스로는 Flume, Scribe, Chukwa 등이 있다. 두 번째 데이터 저장 단계는 크게 원본 데이터 저장과 트랜젝션 데이터 저장으로 나눌 수 있다. 원본 데이터 저장의 경우 수집된 데이터를 안정적으로 저장하는 저장소, 즉 비구조적 데이터 저장소로 주로 대용량 파일 저장소가 이에 해당한다. Hadoop File System, MogileFS가 대표적이라고 할 수 있다. 다음으로 분석 단계에서는 데이터 수집과 동시에 분석을 수행하는 실시간 분석 플랫폼과 전체 또는 부분 데이터에 대해 복잡하고 다양한 분석을 수행하는 배치 분석 플랫폼이 있다. 실시간 분석 플랫폼은 복잡한 분석보다 count, sum등 단순한 aggregation 연산 정도를 수행하는 것이며, S4, Storm 등의 오픈소스 솔루션이 존재한다. 반면 배치 분석 플랫폼의 경우 대용량 처리를 위해 분산, 병렬처리를 필요로 하며 단순 텍스트 분석부터 그래프 분석까지 다양한 분석 모델을 지원한다. Hadoop 분산 처리를 위한 MapReduce를 포함하여 Girahp, GoldenOrb가 있다. 그 밖에도 Cluster, Classification 등과 같이 데이터 마이닝을 위한 데이터 마이팅/통계 도구도 빅데이터를 위한 필수기술에 해당하며 Mahout, R 등이 이에 해당한다. 마지막으로 클러스터 관리 및 모니터링, 데이터 Serialization은 데이터 표현을 위한 기술이다. 대부분 분산 시스템으로 구성되기 때문에 전체 클러스터에 대한 관제 및 모니터링이 복잡해지며, 이를 위해 ZooKeeper, HUE, Cloumon 등이 있다. 또한 이기종 플랫폼 및 다양한 종류의 솔루션을 사용하기 때문에 데이터 전송 및 처리에 대한 표준 프레임워크 또한 필요하며 대표적인 오

픈소스로는 Thrift, Avro, ProtoBuf이다.

7. 빅데이타 적용사례

아직도 초기 기술로서 그다지 많은 기업이 도입하고 있지는 않지만 대기업 또는 인터넷 서비스 기업을 중심으로 시험 도입이 실시되고 있다. 따라서 전사적인 적용보다는 고객 서비스 중심 업무에 적용되는 사례를 중심으로 살펴보기로 하였다.

7.1 제조업체의 수요예측 분석

국내 글로벌 전자제조업에서 고객의 feed back을 SNS의 빅데이타로 실시간으로 수립하는 사례를 들 수 있겠다. 이 회사는 제품의 판매량, 경쟁사 현황, 제품인지도, 노출도등을 dashboard 형태로 구성하고 Trend를 분석하고 있다. 이러한 데이터를 지속적으로 수집하고 분석하다가 갑자기 이상 징후가 나타나면 이에 해당되는 현상에 부합되는 SNS데이타를 기초데이타로 찾아낸다. 이렇게 찾아낸 데이터의 의미를 분석하여 설계나 생산에 반영하는 시스템을 적용하고 있고 판매확대를 예측한다.

이러한 분석 기법은 이제까지 사용하지 않던 소셜네트웍데이타를 활용하는 시도인 것이다.

7.2 이벤트 마케팅 분석

광고라는 홍보수단이 기업에서는 중요한 마케팅 수단이지만 그 효과를 분석하는 것이 쉽지 않은 것은 사실이다. 모든 일이 비용대비효과가 명확해야만 지속적인 시도를 할 수 있다. 따라서 고객의 인지도라는 비정형적인 효과를 밝혀내는 것이 매우 어려운데 빅데이타 분석을 통하여 이러한 분석을 시도하고 있다. A광고회사에서도 유명 탈렌트 별로 광고 후 고객들이 기억하고 있는 기간을 분석하여 광고 효과의 유효성 또는 지속성을 분석하고 있다. 즉, 광고 후 여러 매체에 나타나고 있는 리플을 수집하여 그 양에 따라 효과를 계량하고 있다. 예를 들어, 광고모델로서 걸그룹의 효과가 오래가는지 운동선수의 효과가 오래가는지를 분석할 수 가 있다.

7.3 비정형로그데이타분석을 통한 서비스품질 분석

이동통신회사에서 기지국내의 통화품질을 좌우하는 것은 기지국내의 교환장비에 수집되는 로그데이타에 의해 좌우된다. 그러나 이 로그데이타가 시간당 5천만건이상이 되고 있어서 이를 처리하는데 90분이상이 소요되고 있다. 그러나 통화 품질 장애 판단시 10분내에 조치하여야 하는데 분석에 많은 시간을 소요하므로 업무처리요건 만족도가 매우 낮아 질 수 밖에 없다. 즉 장애가 발생한지 90분이 지나서야 장애로 판정이되어 진다며 가입자의 불만이 증폭될 수 밖에 없다. 이에 Hadoop기반의 분석시스템을 구축하여 2분 29초만에 처리하고 기존대비 요건 불만족도를 97%나 감소하는 효과를 보았다.

7.4 생산설비의 장애분석 시스템

반도체회사 생산공정에는 수많은 장비가 유기적으로 연결되어 생산시스템을 이루고 있다. 600여대의 장비의 장애나 이상상황을 분석하는 업무가 기존에는 2일 이상 걸렸으며 장애가 실시간으로 통보되지 못하고, 또한 장애분석인원 역시 10명에 달했으며 버그분석 및 수정이 월단위로 이루어졌다. 그러나 빅데이타 분석 도구를 활용하고 나서 장애발생 후 5초 이내에 SMS를 통해 통보되며, 분석시간 역시 10분이내에 완수되었다. 장애예측도를 세팅에 의해 예측되고 있으며 분석인원은 1명으로 처리하게 되었다. 따라서 인건비를 절약하고 장애로 인한 비즈니스 손실 또한 획기적으로 절감한 사례이다.

8. 빅데이타에 대한 업계 동향

글로벌 IT 벤더들은 기존에 DW에서 확장된 어플라이언스 제품을 중심으로 대용량급의 H/W을 출시하고 있으며 이에 Hadoop을 적용하여 빅데이타를 분석할 수 있다고 마케팅전략을 펼치고 있다. 오라클은 엑사데이타, EMC는 그린플럼과 ATMOS, IBM은 넷테자와 빅데이타 분석 플랫폼을 제시하고 있으나 아직까지 이렇다할 강자는 나타나지 않고 있다. 솔루션으로서는 Splunk, Datastorm등 구축사례를 중심으로 시장에 제품을 출시하고 있으며, 광범위한 분야에 레퍼런스를 시도하고 있다.

III. 결 론

빅데이타 분석은 이제 시장에 막 진입하는 기술로서 많은 부분이 새롭게 정의되고 해결책이 나타나야한다. 특히 이러한 기술을 활용할 수 있는 전문가는 너무 부족하고 앞으로도 많이 필요할 전망이다. 따라서 데이터 분석능력은 모든 관리자의 기본 역량으로 자리매김할 것이다. 이러한 현상은 Data Scientist라는 직종을 만들게 되었고 그들은 기업 내부의 다양한 영업보고서부터 소셜네트워크의 고객정보까지 방대한 디지털 데이타를 분석, 비즈니스 전략수립을 돕는 역할까지 하게 된다. 이들은 빅데이타가 생성되는 구글등 대형 인터넷 기업뿐만 아니라 유통, 미디어, 심지어 정치분야까지도 필요로 하게 되었다. 최근 미국오바마 선거캠프에서도 소셜네트워크 분석을 위한 빅데이타팀을 꾸렸다고 한다. 이러한 일들을 하기 위해서도 데이터 과학자의 역량이 기술적인 숙련도는 물론이거니와 호기심, 스토리텔링능력, 영리함과 독창적인 시각을 가진 인문학적 능력이 요구되고 있다. 또한 빅데이타 시대의 도입전략으로 첫째, 데이터 접근에 대한 관리가 중요한 보안, 지적재산권, 지적재산권, 개인정보보호 등 법적 책임 관련 준비가 세심하게 관리되어야 하며 외부데이타를 어떻게 내부에서 활용해야 할지 방안을 심도있게 검토해야 한다. 둘째, 여러곳에 나누어져있는 데이터를 클라우드 기반으로 통합하고 데이터 공유 프로세스를 정립함으로서 일관된 관리가 이뤄져야 한다. 셋째, 내부에 있는 데이터베이스와 결합분석을 통한 분석 시스템이 구성되고 실시간 의사결정이 될 수 있도록 지원방법은 마련해야 한다. 마지막으로 앞서 말한 빅데이타 분석을 위한 데이터 과학자 조직을 만들고 인사이트를 끌어 낼 수 있는 전문가를 양성 또는 채용하여야 한다. 초기기술로서 빅데이타 프로젝트는 성공률이 15%미만이지만 지속적인 시행착오와 적극적인 도입전략을 펼친다면 차별적인 경쟁우위를 갖추는 무서운 무기가 될 것이다.

참고문헌

[1] “빅데이터의 가치와 도입전략”, 투이컨설팅, 2012. 02.
[2] “SNS시대의 하이브리드 빅데이터 분석 기술 및 사례”, 솔트룩스, 2012. 02.
[3] “빅 데이터 활용을 위한 빅 애널리틱스 전략”, 위세아이텍, 2012. 02.
[4] “오픈소스를 이용한 빅데이터 라이프 사이클 관리”, SK C&C, 2012. 02.
[5] “Big data개요”, IBM, 2011. 11.
[6] “새로운 시장, 그리고 새로운 기회”, EMC, 2011.
[7] “Cloud, Big data세미나”, EMC, 2012. 01.
[8] “Reshaping Storage infrastructures to Support Virtualization and Big Data Initiatives”, Gartner Symposium, 2011.
[9] “Big Data! -Converging Data Architectures-”, Datastorm, 2012. 02.

저자소개

이 성 희
- 성균관대학교 학사
- 연세대학교 산업공학과 공학석사
- LG엔시스 기술 부문장(CTO)
- 한국CSO협회 운영위원
- 한국클라우드 서비스협회
- 기술분과위원
- U-부천 포럼 상임이사
- IT서비스산업협회
- 가상화포럼기술위원