특집 06

빅데이터를 위한 EDW(Enterprise Data Warehouse) 고도화 방안

유민석 (현대정보기술)

목 차 »

1. 서 론

빅데이터는 최근 가장 주목받고 있는 IT 패러다임으로서 기업들의 고객 데이터 수집활동 및 멀티미디어 콘텐츠의 폭발적인 증가와 스마트폰 보급, SNS 활성화 및 사물통신망의 저변 확대로 빠르게 확산되고 있다.

데이터 용량에 대한 전망은 그 동안 여러 번 언급되었다. 2020년까지 디지털화되어 저장될 데이터의 용량은 35조 GB[1]에 달해 2009년의 44배 수준이 된다고 한다. IDC의 조사에 따르면, 이미 2010년 말에 120만 PB, 즉 1.2 ZB 수준의 데이터 용량을 달성했다. 이를 DVD에 저장하면 지구와 달 사이를 왕복할 만큼 쌓을 수 있다. 편도로 약 38만 km의 길이다. 데이터가 폭증하면서 가장 발빠르게 대응한 IT분야는 구글(Google), 페이스북(Facebook), 야후(Yahoo), 링크드인(Linkedin)등과 같은 대고객 서비스를 제공하는 업체들이었다. 이들은 방대한 데이터를 활용하는 방법으로 오픈 소스 프레임워크 기반의 하둡(Hadoop)을 활용하였다. 하둡(Hadoop)[2]은 대량의 자료를 처리할 수 있는 서버 클러스터에서 동작하는 분산 응용 프로그램을 지원하는 자유 자바 소프트웨어 프레임워크이다.

초기의 빅데이터 기술은 SNS 및 텍스트 기반의 대량 데이터를 더 저렴하고 빠르게 데이터를 조작하고 분석하는 업무에 적용되었으나, 향후

1) <참고:디지털 정보 단위>

B(byte)	KB	MB	GB	TB	PB	EB	ZB
바이트	킬로 바이트	메가 바이트	기가 바이트	테라 바이트	페타 바이트	엑사 바이트	제타 바이트
	1024B	1024KB	1024MB	1024GB	1024TB	1024PB	1024EB

2) 하둡(Hadoop)은 대량의 자료를 처리할 수 있는 큰 컴퓨터 클러스터에서 동작하는 분산 응용 프로그램을 지원하는 자유 자바 소프트웨어 프레임워크이다. 원래 너치의 분산처리를 지원하기 위해 개발된 것으로, 아파치 루씬의 하부 프로젝트이다. 분산처리 시스템인 구글 파일 시스템을 대체할 수 있는 하둡 분산 파일 시스템(HDFS: Hadoop Distributed File System)과 맵리듀스를 구현한 것이다.

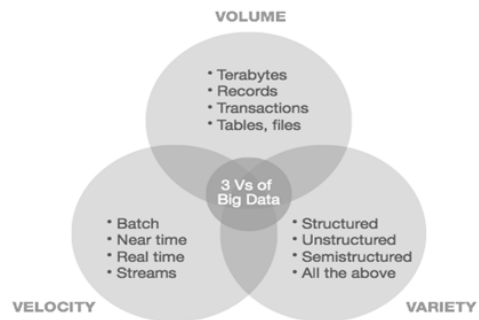
슈퍼컴퓨팅이 확산되면 빅데이터 기술은 빠르게 전파돼 기업의 주요 업무 처리 방식을 바꿔 놓을 것으로 기대된다. 본고에서는 현재 가장 크게 주목받는 빅데이터 구현 환경인 하둡을 기반으로 EDW 고도화 방안을 제시하도록 한다.

2. 기반 기술

시스템 비용이 낮아지고 처리 능력은 향상되면서 빅데이터가 주목 받기 시작했다. 메인 메모리 가격이 내려감에 따라 기업들은 그 어느 때보다도 많은 데이터를 메모리 내에서 처리할 수 있게 됐다. 또한 서버 클러스터에 컴퓨터들을 연결하기가 더 쉬워졌다. 과거에는 일부 대형 슈퍼컴퓨터만이 다른 시스템과 연동돼 다중으로 처리할 수 있었다. 이 슈퍼컴퓨터들은 특화된 하드웨어였기 때문에 가격이 수십만 달러 이상이었다. 이제는 상용화된 하드웨어를 이용해 그런 가능성을 달성할 수 있다. 이를 통해 우리는 데이터를 더 빠르면서 저렴하게 처리할 수 있게 되었다.

IDC의 DBMS 담당 애널리스트 칼 올롭슨[3]은 3가지(3V)가 조합돼 빅데이터가 생겨났다고 말했다. 3V는 다양성(Variety), 볼륨(Volume), 속도(Velocity)를 의미한다. 다양성은 구조화된 데이터뿐 아니라 비정형 데이터 형태로도 들어온다는 의미다. 볼륨은 수집되고 분석되는 데이터의 양이 매우 크다는 것을 의미한다. 그리고 속도는 데이터가 처리되는 속도를 뜻한다. 현재 3V를 커버할 수 있는 기술로 가장 크게 주목받는 것은 하둡이다. (그림 1)은 3V의 특성을 보여준다.

하둡은 아파치 재단의 오픈소스 프로젝트로서 신뢰성 있는 대규모 분산 컴퓨팅을 가능하게 하는 플랫폼이다. 하둡은 슈퍼컴퓨팅에서 가장 보편적인 접근방식인 맵리듀스를 기반으로 하고 있



(그림 1) 3V의 특성[3)]

으면서도 구글이 지원하는 프로젝트를 통해 단순하면서도 고급스럽게 변화되었다. 이 장에서는 하둡을 구성하는 주요 프로젝트인 하둡 분산 파일시스템(HDFS), 맵리듀스(MapReduce), 하둡 데이터베이스(HBase) 및 데이터웨어하우스 인프라스트럭처를 지원하는 하이브(Hive)에 대하여 기술한다.[4]

2.1 HDFS(Hadoop Distributed File System)

대용량 데이터를 저장할 수 있는 분산 파일 시스템으로 수천대 규모의 저가 서버 클러스터를 묶어 단일 파일 시스템 이미지를 제공하여 비용 절감 효과와 함께 뛰어난 확장성을 보장한다. 특히 데이터 안정성을 보장하기 위해 최소 세 개의 복사본을 유지하며 대용량을 커버하기 위해 64 MB의 큰 블록 단위를 가지고 있는 것이 특징이다.[1,5]

2.2 맵리듀스(MapReduce)

분산 데이터 처리 시스템으로 HDFS에 분산 저장되어 있는 데이터를 map()과 reduce()라는 간

3) 참고:[3] 및 TDWI Research 2011 Big Data Analytic Report

단한 분산 프로그래밍 방식을 통해 병렬 처리 해 준다. 분산 병렬 처리에 필요한 작업 스케줄링, 부하 분산, 장애 대책 등을 시스템에서 처리해 주기 때문에 쉽게 data parallel 스타일의 병렬 처리를 가능케 한다.[1,5]

2.3 HBase(Columnar NoSQL Store)

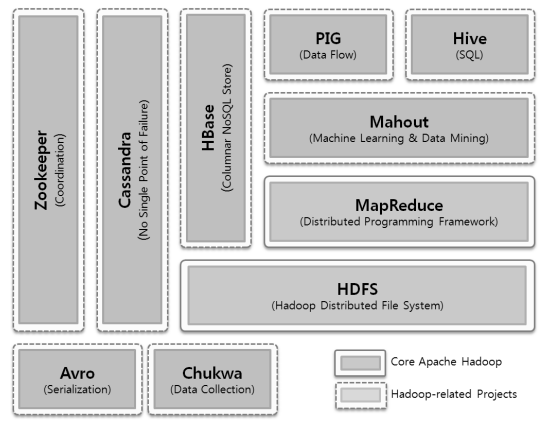
HDFS 기반의 분산 데이터 저장소로서 기존 관계형 데이터베이스와 달리 컬럼 기반의 Key-Value 방식의 저장방식을 채택하고 있으며 대량 데이터의 확장성 및 분산 환경을 보장하는 정형 데이터 저장소 역할을 수행한다. HBase의 성능이 지속적으로 개선되고 있으나 기존 SQL 데이터베이스의 완전한 대체보다는 페이스북의 메시지 플랫폼 등 데이터 중심의 웹사이트에 활용되고 있다.[1,5,6]

2.4 하이브(Hive)

데이터 요약, 조회 및 분석을 위한 데이터웨어하우스(DW) 인프라 역할을 수행한다. 초기 페이스북에서 개발된 스펙을 기반으로 기능이 향상되었고, 주요 특징으로는 Amazone S3와 같이 하둡(Haoop) 분산 파일 시스템 호환성을 유지하는 대용량 데이터에 대한 분석을 지원하고, 기존 RDBMS SQL과 유사한 HiveQL을 지원한다. [1,6]

3. 설계 및 구현

본 장에서는 RDBMS/DW하에 운영되는 기존 업무 시스템을 하둡기반 빅데이터 플랫폼을 활용한 Migration 방안에 대하여 기술하도록 한다. (그림 2)는 하둡 프레임워크를 도식화한 것이다.



(그림 2) 하둡(Hadoop) 프레임워크[1]

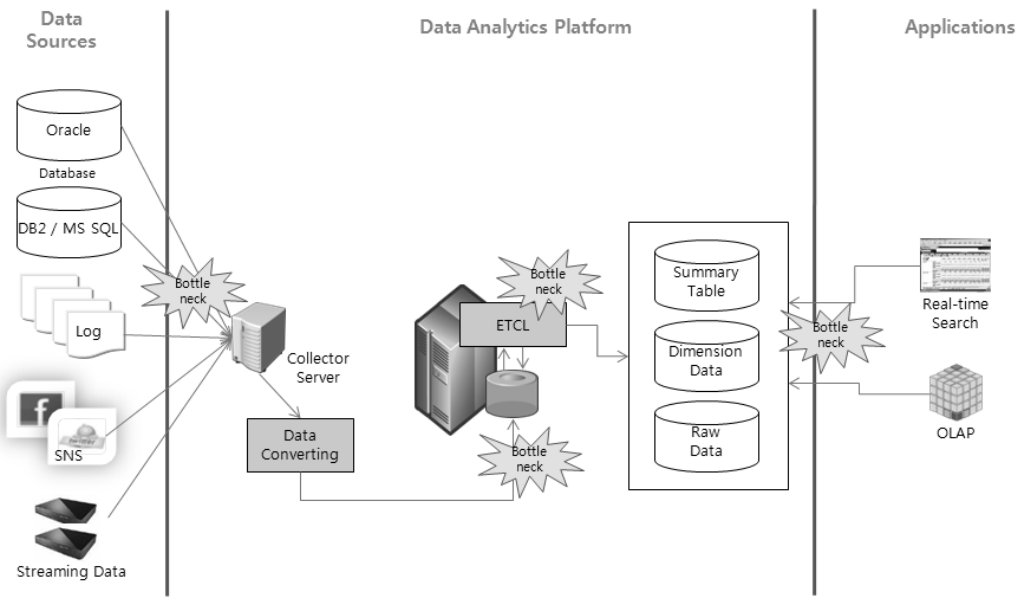
3.1 하둡기반 EDW(Enterprise Data Warehouse) 설계

3.1.1 현행 시스템의 주요 병목 현상 파악

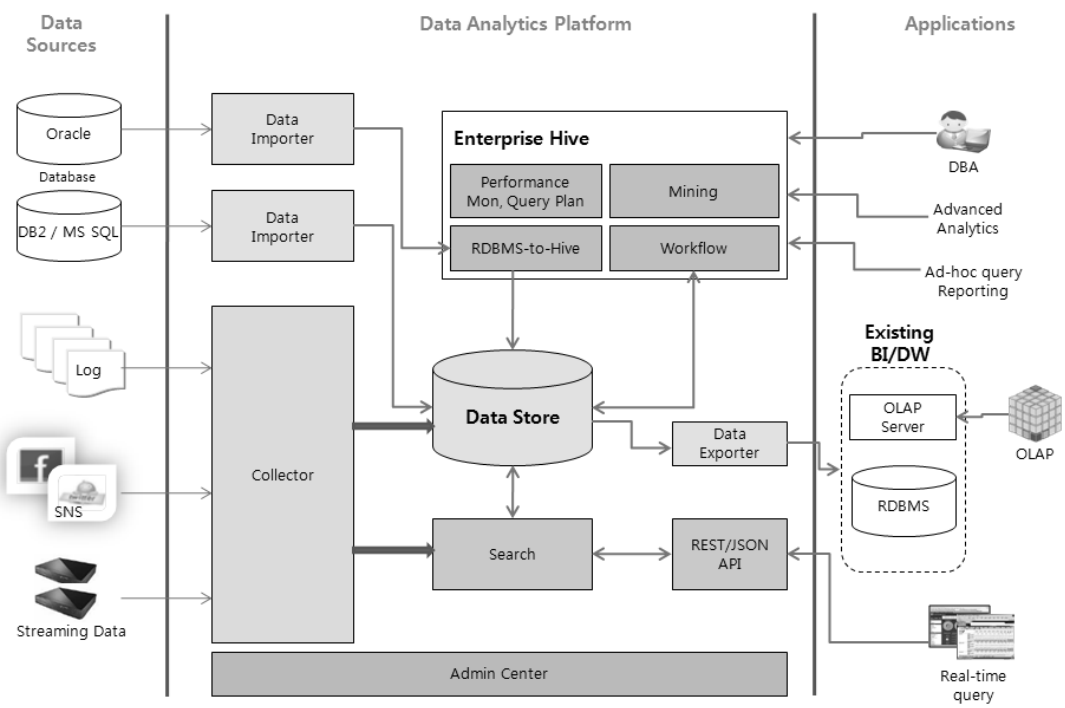
(그림 3)은 일반적인 EDW기반 분석시스템의 주요 병목 점을 보여준다. 첫 번째는 분석에 필요한 원천 데이터의 취합 부분에서 병목 현상이 발생하고, 두 번째는 데이터웨어하우스(DW)및 데이터마트(DM)에 추출/변경/정제 및 적재 시에 병목이 발생하고, 마지막으로 사용자 화면에 필요한 데이터를 표출할 때 지연 현상이 발생한다.

3.2 개선방향을 고려한 상세 구현 설계

기존 시스템의 성능저하에 대한 개선방향, 멀티미디어 콘텐츠의 폭발적인 증가와 스마트폰 보급, SNS 활성화 및 사물통신망의 저변 확대로 빠르게 늘어나는 데이터 볼륨에 대한 개선 방향 및 신규 비즈니스 서비스에 대응하기 위한 정형 및 비정형 데이터를 처리할 수 있는 개선방향을 고려하여 상세 구현 방안을 제시한다. (그림 4)는 3V(다양성(Variety), 볼륨(Volume), 속도(Velocity)) 문제점을 개선할 수 있는 좋은 예로 판단된다.



(그림 3) EDW기반 분석시스템의 주요 병목점



(그림 4) 하둡기반 EDW 고도화 방안[8,9]

4. 구현 사례 비교분석

본 장에서는 하둡기반 빅데이터를 위한 EDW 고도화가 이루어진 국내 대표적인 통신회사인 A사의 사례를 집중적으로 살펴보고, 추가적으로 빅데이터를 활용한 다양한 응용 사례를 제시하고자 한다.

4.1 A사(국내 통신사) CDR(Call Detail Record) 사례[7]

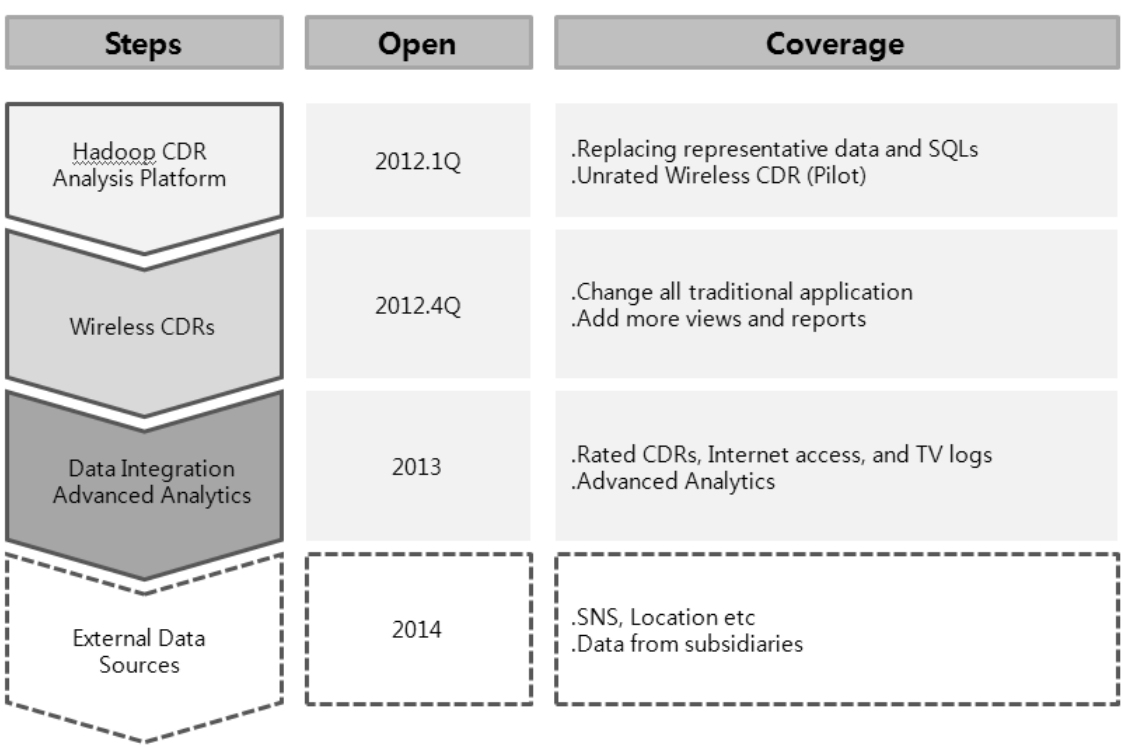
4.1.1 추진 배경 및 전략

확장성 측면에서는 Wired, 2G, 3G, WiMax, LTE, WiFi, SMS, MMS등의 다양한 데이터 유형 및 증가하는 데이터양에 대한 대응 방안이 필요하였고, 데이터 수집, 데이터 저장 스토리지 및 데이터 검색 등에 대한 수평적 확장성 보장이 필요하였다. 성능 측면에서는 실시간에 준하는 스트리밍 CDR 데이터 분석 요구가 있었다. 마지막으로 상용화된 하드웨어 클러스터를 활용하여 비용 효율성 측면을 고려하여 진행하였다.

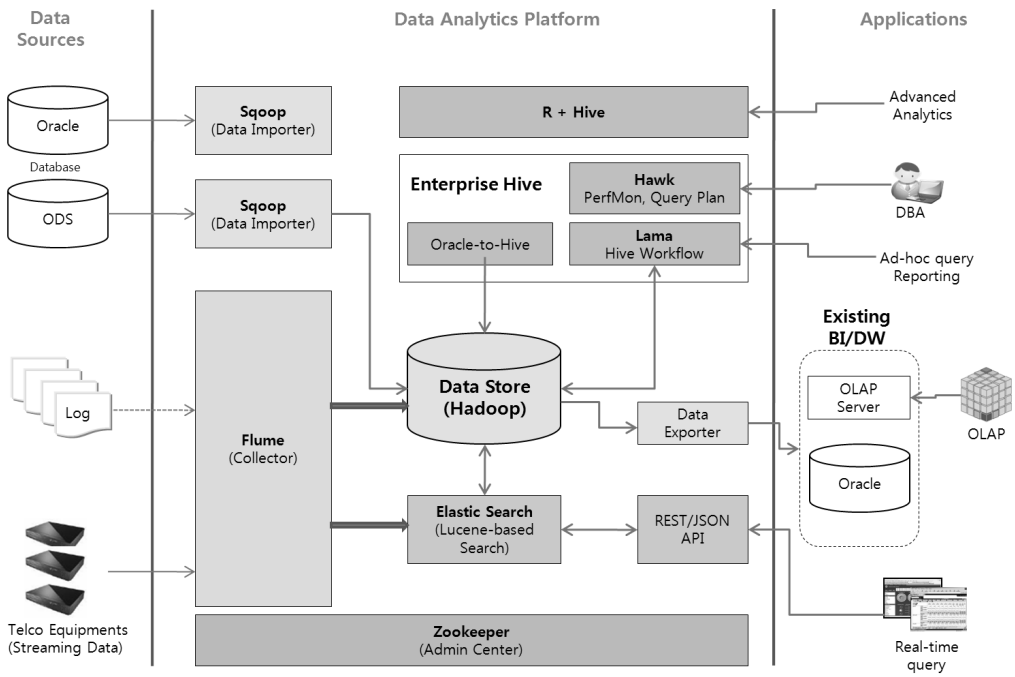
A사의 구현 전략은 신기술 적용이라는 현실적 제약 사항을 반영하여 빅뱅(Big Bang) 방식보다는 점진적 구현 방식(Step by Step)을 채택하였고, (그림 5)는 A사의 점진적 구현 일정을 보여준다.

4.1.2 플랫폼 구조

(그림 6)은 A사 CDR 시스템에 적용된 플랫폼 구조를 보여준다.



(그림 5) A사 점진적 구현 일정



(그림 6) A사 적용 플랫폼 구조

4.1.3 시사점

신규 기술을 적용하는 관계로 업무에 맞는 올바른 솔루션을 선택하는 것이 첫 번째 스텝이다. 오픈 소스 프로젝트 속성상 여러 개의 하위 프로젝트별 장점을 취합하여 병합할 필요성이 발생하고, 특정 업무 요청 사항이 필수 불가결한 경우에 오픈 소스를 수정하여 반영할 수 있는 능력을 보유해야 한다.

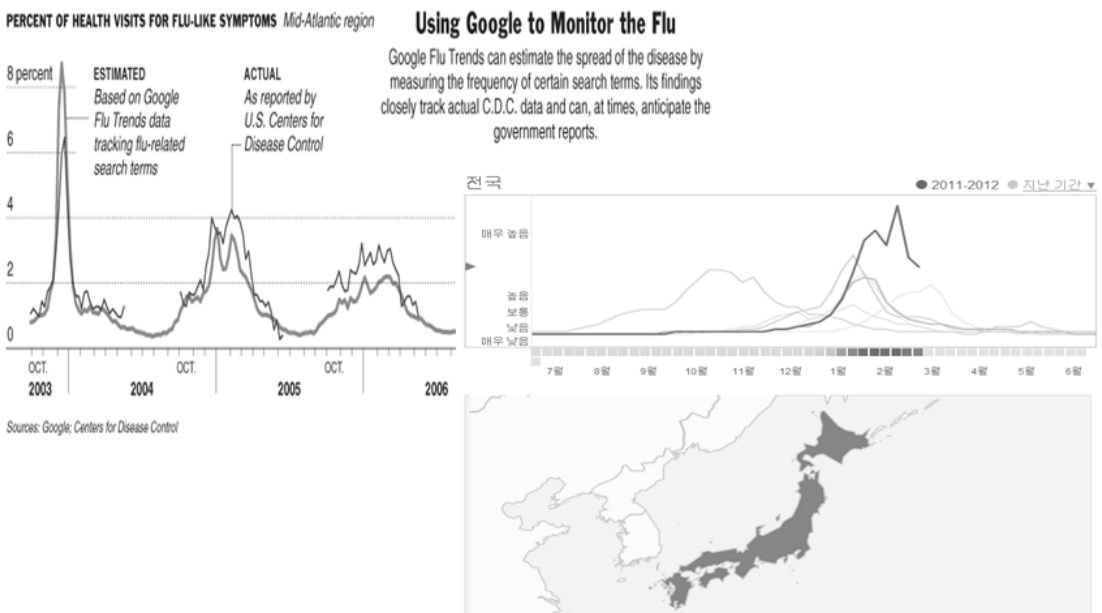
다양한 데이터 유형에 대한 통합이 점점 중요해지고 있다. 특히 정형적인 데이터를 주로 취급하는 DW는 향상된 분석 능력을 보유하기 위하여 정형 및 비정형 데이터를 통합하는 방안을 수립하는 것이 중요하다.

4.2 기타 빅데이터 적용 사례

빅데이터 적용의 대표적인 사례는 구글의 독감 동향 서비스이다. 구글은 미국 질병통제예방센터(CDC)의 예측보다 더 빠르고 정확하게 독감 동향을 진단할 수 있다. 전 세계 어느 지역에서 독감이 유행하며, 또한 유행할 것인지에 대한 분석 서비스를 제공하고 있다. 분석 원리는 사람들이 감기증세를 느끼면 독감과 관련된 키워드를 검색하게 된다. 구글은 이러한 검색의 빈도를 지역별로 분석하여 독감의 확산경로를 추적할 수 있다.

이러한 분석은 구글이 가진 대용량의 검색 데이터와 분석기반을 통해 가능한 것이다. (그림 7)은 구글의 독감 동향 서비스를 보여준다. 본 서비스를 통하여 극동아시아 중 일본에 독감이 3월까지 유행하고 점진적으로 감소하리라는 것을 예측할 수 있다.

볼보자동차는 신규 출시 모델의 초기 불량을 파악하기 위하여 약 5만대 이상의 차량을 생산한 후에 초기 불량을 알아낼 수 있었으나, 현재는 차



(그림 7) 구글(Google) 독감 동향 예측 서비스

량에 센서를 부착해 센서 데이터를 분석함으로써 약 1,800~2,000대를 생산할 때 초기 불량을 파악할 수 있다.

이외에도 의료 현장에서는 환자들의 병력이력을 분석해 비슷한 증상의 환자들에게 적합한 치료법을 제공할 수 있고, 금융계에서는 대출 정보 및 정책 자료를 통합하여 분석한 후 고객에게 최적화된 상품이나 서비스를 추천할 수 있다. 빅데이터 활용은 점점 다양한 산업분야에서 활용 영역을 넓혀가고 있다.

5. 결 론

많은 기업들은 대용량 데이터를 관계형데이터베이스(RDB)로 관리했지만 엄청난 데이터의 크기, 정형 및 비정형 데이터 분석, 빠른 처리 속도의 3가지 이슈로 빅데이터가 부상했다. 빅데이터는 기업에게 위기가 되기도 하고 새로운 기회를 부여하기도 한다. 전문가들은 빅데이터를 도입하기 위해서는 툴이나 솔루션이 아니라 통찰력을 얻기 위한 기업내부의 데이터를 먼저 찾아야 한다고 얘기한다. 비즈니스에서 활용하기 위한 목적으로 접근해야 한다는 것이다. 즉, 어디서 어떤 정보를 가져올 것이며 어떤 기반에서 분석을 수행할 것인지 판단해야 한다.

기업들은 빅데이터를 활용하는데 있어서 다음 사항을 고려해야 한다. 빅데이터 분석이 산업군별로 다양한 만큼 서비스나 미래의 비즈니스를 위해서 현재 활용하고 있지 못하는 데이터가 무엇인지 찾고, 기존에 어떤 사례가 있는지 검토한 후 단계별 검증을 통해서 비즈니스에 반영하는 것이 필요하다.

앞서 살펴 본 것처럼 이제는 빅데이터를 감당할 수 있을 뿐 아니라 적절하게 처리할 수 있는 수단이 마련되었다. 정형 및 비정형 데이터를 아우르는 빅데이터 기반 EDW 사례가 현재까지 많지는 않으나 A사 및 기타 사례처럼 기술적으로 뛰어난 하둡 플랫폼은 시장에서 인정받고 많은

사용자 및 개발자를 확보하여 점점 더 경쟁력 있는 플랫폼으로 성장하리라 판단된다.

참 고 문 헌

[1] Hadoop, http://hadoop.apache.org
[2] 권갑현, 클라우드 컴퓨팅을 위한 하둡 작업 스케쥴링 알고리즘의 비교 연구, 동양대학교 논문집, Vol.14 No.1, 2010.
[3] IDC, 칼 W. 올롭슨, 빅데이터의 중요성(The Big Deal about Big Data), 2011년 2월.
[4] 톰 화이트 저, Hadoop 완벽 가이드, 한빛미디어, 2011년 5월.
[5] 한재선, 클라우드 컴퓨팅 플랫폼과 오픈 플랫폼 기술, 정리처리학회지 제 16권 제2호, 2009년 3월.
[6] Wikipedia, http://en.wikipedia.org/wiki/
[7] 구자형, 한재선, Replacing RDB/DW with Hadoop and Hive for Telco Big Data, 2011년 11월.
[8] 에벤 휴잇 저, 송무찬/최원우 편역, Cassandra 완벽 가이드, 한빛미디어, 2011년.
[9] 채승병, 빅데이터(Big Data) 분석과 활용, 삼성경제연구소, 2011년 2월.
[10] KRG, 2012년 IT시장전망 세미나 자료집, 2012년 2월.
[11] 오석균, 송민구, BI 가이드 3.0, 현대정보기술, 2012년 2월.

저 자 약 력

유 민 석

이메일 : msyoo@lotte.net

- 1989년 광운대학교 전자계산학과(학사)
- 2011년~현재 현대정보기술 컨설팅사업부 / BI팀장
- 관심분야 : 전사데이터웨어하우스(EDW), 데이터모델링, 고객관계관리(CRM), 콜센터(Call Center) 응용