DW 어플라이언스를 통한 빅데이터 처리 기술 동향 분석

최로환*, 박석천**, 심봉수***
*가천대학교 일반대학원 모바일소프트웨어학과
**가천대학교 컴퓨터공학과 정교수(교신저자)
***데이터스트림즈 기술연구소 책임연구원
romance213@gc.gachon.ac.kr

Analysis of Trend for BigData Processing Technology by DW Appliance

Ro-Hwan Choi*, Seok-Cheon Park**, Bong-Soo Sim***
*Dept of Mobile Software, Gachon University
**Dept of Computer Engineering, Gachon University(Corresponding Author)
***Senior Researcher, DATASTREAMS co., ltd

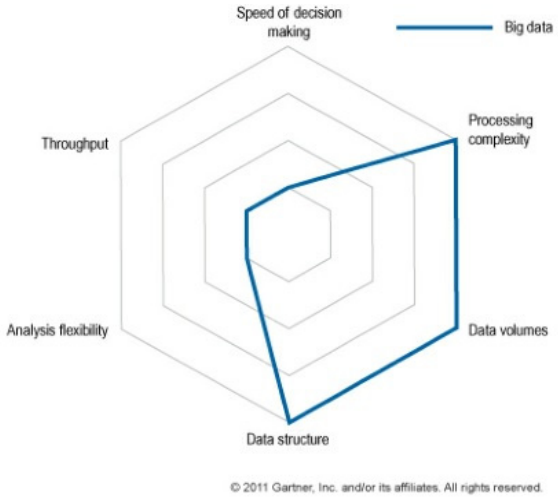
요 약

최근 정보통신기술이 하루가 다르게 발전함에 따라 하루에도 수많은 데이터가 흘러나오는 최근의 추세이다. 정형 데이터 뿐 아니라 비정형 데이터 분석까지 진행하는 최근의 추세에 맞춰 현 빅데이터 기술 동향을 분석한다. 빅데이터 시대를 맞아 기존의 데이터웨어하우스(DW)와 발전된 데이터웨어하우스(DW) 어플라이언스에 대해 분석하고 향후 발전 전망과 방향을 제시한다.

Ⅰ. 서 론

최근 정보통신기술이 하루가 다르게 발전함에 따라 하루에도 수많은 데이터가 흘러나오는 것이 최근의 추세이다. 특히 스마트폰과 모바일 인터넷 서비스 활성화에 따라 사용자의 이용데이터가 폭발적으로 증가하게 되었다. 이렇게 데이터가 폭발적으로 증가한지는 불과 얼마 되지 않았지만 그동안 기하급수적으로 늘어난 데이터양은 모바일을 포함한 다양한 플랫폼에서도 하루에도 수 테라바이트의 빅데이터를 생성해내고 있다.

다음 그림 1은 Gartner에서 제시한 빅데이터의 성격을 보여주는 그래프이다.



(그림1) 빅데이터의 성격 그래프

빅데이터는 부피가 크고(Volume), 변화의 속도가 빠르며(Velocity), 데이터의 속성이 너무도 다양한 데이터를 지칭하는 것(Variety)이라고 정의하고 있다.

산업계에는 이 빅데이터를 바라보는 시각으로 크게 두 가지로 나눈다. 하나는 비정형 데이터를 중심으로 효과적으로 데이터를 처리하는 기술과 정형 데이터베이스에서 대규모 저장 시스템을 연구하는 분야이다. 현 업계에서 이슈가 되고 있는 것은 바로 전자에 속하는 분야이며, 후자의 빅데이터는 기존의 대용량 데이터베이스와 장비 시장을 장악하고 있는 글로벌 기업들을 중심으로 부각되고 있는 빅데이터를 말한다[1].

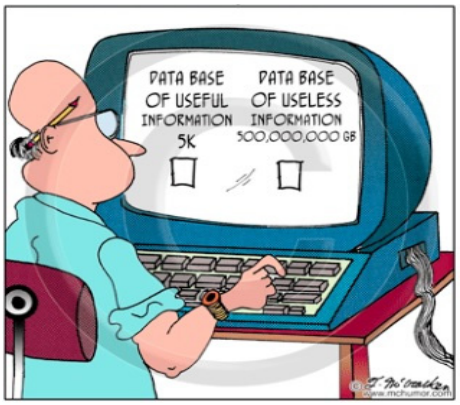
바로 이 전자의 분야, 즉 비정형 데이터를 중심으로 데이터를 처리하는 기술들은 그 동안, 저장장치의 단가절감의 영향으로 인해 잘 가공된 데이터와 버려지는 데이터를 나누었었다. 기존의 버려지고 있는 분석할 가치가 있을지 없을지도 모르는 상황 때문에 큰 비용을 들여 처리하는 것은 비효율적이었기 때문이다. 하지만 최근 대상이 되는 데이터를 전부 처리하는 것이 아니라 필터링 및 클러스터링을 통해 데이터를 효과적으로 걸러내는 기술이 등장함에 따라 잘 가공된 데이터가 아닌 그 동안 버려지는 데이터의 활용 가능성까지 데이터마이닝 할 수 있게 되었다.

그런데 이러한 비정형 데이터를 처리 가공하기 위해서는 여러 가지 어려움이 따른다. 먼저 처리 복잡도가 높고 처리할 데이터양이 광범위하며 동시 처리량이 낮다[2].

이러한 비정형 데이터와 가공되지 않은 정보가 난무하

는 가운데 DW(Data WareHouse) 어플라이언스의 중요성이 커지고 있는 실정이다.

다음의 그림 2는 실제 유용한 데이터는 많은 데이터 가운데 소수에 불과하다는 것을 나타낸다.



(그림 2) 전체 데이터 중 유용한 소스 예

Ⅱ. 관련 연구

2.1 데이터웨어하우스(DW) 어플라이언스의 등장

기존 DW의 개발 과정은 매우 복잡한 과정을 거쳐 이루어진다. 먼저 질의 처리 및 결과의 최적 경로 비용을 찾아내는 옵티마이저를 세팅하고, DW 프로젝트의 성패를 가름하는 물리 및 논리 모델링 작업을 진행해야 한다. 이외에도 상당 기간의 테스트 과정과 시간을 들여 최적화된 인덱스를 설정하는 것을 비롯해, 업무에 적합한 마트 설계 등의 과정을 거쳐야 한다.

이런 복잡한 과정을 거쳐도 종래의 방식은 업무가 정형화돼 있을 때만 DW 프로젝트의 성공을 보장할 수 있다는 단점을 갖고 있다. 그러나 실제 현업의 요구 사항은 규정된 범위 내에 존재 하지 않는 경우가 많다. 또한 개발 기간에는 도출되지 않았던 문제점들이 개발 이후 유지 보수 과정에서 나타날 수도 있다.

이러한 DW의 성능 이슈에 대한 문제점을 해결하기 위해 등장한 개념이 바로 DW 어플라이언스다. DW 어플라이언스는 데이터베이스관리시스템(이하 DBMS), 서버, 스토리지를 구조적으로 통합한 것으로, 성능을 최대화 하도록 하드웨어와 소프트웨어를 최적화 한 것이다. DW 기능에 맞춰 스토리지와 DBMS를 최적화했기 때문에 서버, 스토리지, DBMS를 별도로 도입해 DW를 구축하던 기존의 방식과 달리 튜닝 작업이 크게 단축되고 성능이 향상됐다. 즉, DW 어플라이언스는 DBMS와 서버, 스토리지를 하나로 묶어 DW를 구현하는 형식을 취하고 있다.

2.2 데이터웨어하우스(DW) 어플라이언스 구성

DW 어플라이언스는 하드웨어 영역인 서버와 스토리지, 그리고 네트워크와 함께 소프트웨어 영역인 DB 분석 솔루션을 포함한다. DB 분석 솔루션은 데이터웨어하우스(DW)의 생성과 관리, 비즈니스 인텔리전스(BI), 온라인 분석 프로세스(OLAP), 데이터마이닝, 분석 등 사용자 의사 결정을 위한 툴을 제공하는 솔루션을 총칭한다. 방대한 데이터 속에서 전문적이고도 정확한 분석을 가능하게 하는 S/W 지원 영역이라고 볼 수 있다.

다음의 그림 3은 데이터웨어하우스(DW) 어플라이언스의 구성을 나타낸다.

(그림 3) 데이터웨어하우스 어플라이언스 구성

2.3 데이터웨어하우스(DW) 어플라이언스 기술

데이터웨어하우스(DW)는 시스템 혹은 어플리케이션에 분산되고 단절되어 있는 정보를 추출(Extract)/변환(Trans-formation)/통합(Integration)하는 과정을 거쳐 적재 및 저장(Load)함으로써, 사용자가 원하는 분석정보를 제공하는 환경을 의미한다. 일반적으로 서버 및 스토리지 같은 H/W와 데이터베이스 분석 솔루션으로 구성된다.

최근 데이터웨어하우스(DW)의 용량은 급격하게 커지고 있다. 그러한 이유는 데이터로 인한 수익발생이 생긴 뒤 그것에 따른 플랫폼도 함께 협업이 가능하게 되어 동일한 데이터로 하여금 활용도가 높아지게 되었고 그것에 따른 추가 요구사항의 확립으로 더욱 효율적인 데이터웨어하우징이 되어 그러한 연쇄 선기능이 작용하였기 때문이다[3].

데이터웨어하우스(DW)는 빅데이터를 처리하기 위해서 DBMS에게 의존도가 높을 수밖에 없는 기술적인 특성 때문에 단순명료해야 한다.

하지만 단순하게 데이터웨어하우스를 구축하는 것은 쉽지 않다. 그 이유는 DW 구축에 반드시 필요한 ETL 작업이 시스템 성능을 큰 폭으로 저하시키는 문제를 지니기 때문이다. ETL작업은 다양한 소스시스템(Source System)으로부터 필요한 데이터를 추출(Extract)하여 사용자의 요건에 맞게 변환(Transformation)작업을 거친 후 타겟시스템(Target System)으로 전송 및 로딩(Loading)하는 모든 과정을 말한다[4].

본 논문에서는 늘어나는 데이터관리 및 분석을 위한 어플라이언스와 데이터웨어하우스(DW) 어플라이언스의 현황을 분석하고 앞으로의 전망을 예측하여 DW어플라이언스의 새로운 방향을 제안한다.

Ⅲ. 빅데이터 분석과 데이터웨어하우스(DW) 어플라이언스의 기술동향

3.1 빅데이터 분석 어플라이언스 기술동향

빅데이터 분석을 위한 어플라이언스는 다음과 같다. 즉,

오픈소스 배포판 아파치 하둡(ApacheTM HadoopTM), 오라클 NoSQL DB, 하둡을 위한 오라클 데이터 통합 애플리케이션 아답터(Oracle Data Integrator Application Adapter for Hadoop), 하둡을 위한 오라클 로더(Oracle Loader for Hadoop), 오픈소스 배포판 'R' 등을 포함하고 있는 어플라이언스 시스템이다.

3.1.1 Hadoop

하둡은 오픈소스(OpenSource) 분산처리기술 프로젝트로 현재 정형/비정형 빅데이터 분석에 가장 선호되는 솔루션이라고 할 수 있다. 실제로 야후와 페이스북 등에 사용되고 있으며, 채택하는 회사가 늘어나고 있다. 주요 구성요소로 하둡분산 파일 시스템인 HDFS, Hbase, MapReduce가 포함된다. HDFS와 Hbase는 각각 구글의 파일 시스템인 GFS와 빅 테이블(Big Table)의 영향을 받았다. 기본적으로 비용효율적인 x86서버로 가상화된 대형 스토리지(HDFS)를 구성하고, HDFS에 저장된 거대한 데이터셋을 간편하게 분석처리 할 수 있는 Java 기반의 MapReduce 프레임워크를 제공한다. 이외의 Hadoop을 기반으로 한 다양한 오픈소스 분산처리 프로젝트가 존재한다.

다음의 그림 4는 하둡의 구조와 그에 따른 구글의 분산처리 기술을 나타낸다.

(그림 4) 하둡 구조와 구글의 분산처리 기술

3.1.2 NoSQL

NoSQL은 Not-Only SQL, 혹은 No SQL을 의미하며, 전통적인 관계형 데이터베이스(RDBMS)와 다르게 설계된 비관계형 데이터베이스를 의미한다. 대표적은 NoSQL 솔루션으로는 Cassandra, Hbase, MongDB 등이 존재한다. NoSQL은 테이블 스키마(Table Schma)가 고정되지 않고, 테이블 간 조인 연산을 지원하지 않으며, 수평적 확장이 용이하다는 특징을 가진다. 관계형 데이터베이스의 경우, 일관성(Consistency-모든 노드는 같은 시간에 같은 데이터를 보여줘야 한다)과 유효성(Availablilty-일부 노드가 다운되어도 다른 노드에 영향을 주지 않아야 한다)에 중점을 두고있는 반면, NoSQL 기술은 분산가능성(네트워크 전송 중 일부 데이터를 손실하더라도 시스템은 정상 동작을 해야한다)에 중점을 두고 일관성과 유효성은 보장하지 않는다. 이것은 일관성, 유효성, 분산가능성 중 2가지만 보장이 가능하다는 분산 데이터베이스 시스템 분야의 CAP 이론에 따른 것이다. 따라서 대규모의 유연한 데이터 처리를 위해서는 NoSQL 기술이 적합하지만, 안정성이 중요한 시스템에서는 오랫동안 검증된 관계형 데이터베이스를 채택할 필요가 있다.

3.1.3 R

오픈소스 프로젝트 R은 통계계산 및 시각화를 위한 언어 및 개발환경을 제공하며, R언어와 개발환경을 통해 기본적인 통계기법부터 모델링, 최신 데이터 마이닝 기법까지 구현/개선이 가능하다. 이렇게 구현한 결과는 그래프 등으로 시각화할 수 있으며, Java나 C, Python 등의 다른 프로그래밍 언어와 연결도 용이하다. Mac OS, 리눅스/유닉스, 윈도우 등의 대부분의 컴퓨팅 환경을 지원하는 것도 장점이다. 위의 장점들로 인해 R은 통계분석 분야에서 인지도를 높여왔으며, 하둡 환경상에서 분산처리를 지원하는 라이브러리 덕분에 구글, 페이스북, 아마존 등의 빅데이터 분석이 필요한 기업에서 대용량 데이터 통계분석 및 데이터 마이닝을 위해 널리 사용되고 있다.

3.2 데이터웨어하우스(DW)의 기술동향

3.2.1 오라클

오라클은 SUN Microsystems를 인수해 자사의 데이터베이스와 SUN의 하드웨어를 결합한 형태의 DB 어플라이언스를 출시하고 있다. 오라클의 DB 어플라이언스는 Server Node간 섀시의 내부연결(Internal Chassis Wiring)을 통해 우발적으로 케이블이 뽑힘으로 인한 잠재적인 장애 가능성을 낮추고, CPU 코어를 활성화함으로써 비즈니스 요구 증가에 따른 서버 성능의 향상이 가능하고, 필요한 만큼의 라이센스 초기 구매 후 필요에 따라 추가 구매가 가능하다. 4TB 규모의 DB 시스템으로, 노후화된 장비에 대한 빠르고, 효율적인 교체가 필요한 시스템 또는 DB에 대한 HA 구성이 요구되는 시스템에 적용이 가능한 구조를 채택하고 있다.

오라클 데이터베이스 11g(Oracle Database 11g), 오라클 엑사데이터 데이터베이스 머신(Oracle Exadata Database Machine), 오라클 엑사리틱스 BI 머신(Oracle Exalytics Business Intelligence Machine)과 쉽게 통합이 가능하며, 엔터프라이즈급의 성능, 가용성, 지원, 보안 등을 제공하면서 동시에 모든 형태의 데이터를 분석할 수 있도록 설계하였다[5].

3.2.2 EMC

EMC는 DW 업체 그린플럼을 인수한 뒤 2011년 9월 비정형 데이터 저장을 위해 하둡을 탑재한 그린플럼DCA를 출시했다. 이 장비가 등장하기 전까지만 해도 시장은 정형 데이터와 비정형 데이터를 나눠 따로 분석했다. EMC는 맵R을 바탕으로 따로 하둡을 만든 다음 이를 자사 관계형 DBMS와 하나로 묶어 내는 DW를 만들어 내었다.

3.2.3 SAP

SAP 인메모리 어플라이언스(SAP HANA)는 대용량 데이터를 비즈니스 트랜잭션이 발생한 순간 원하는 비즈니스 인사이트로 검색 및 분석할 수 있는 실시간 비즈니스를 실현하는 솔루션으로, SAP HANA를 구성하는 핵심 엔진으로 칼럼(Column), 로우(Row), 오브젝트 등 3가지 저장방식을 메모리나 디스크로 관리하는 하이브리드 데이터베이스로, 대용량병렬처리(MPP) 엔진을 기반으로 데이터 압축기술과 연산엔진을 탑재해 수십억 건의 데이터를 1~2초 이내에 처리하는 DB 엔진을 포함한다[6].

3.3 국내 데이터웨어하우스 어플라이언스 전망

기업은 보유하고 있는 고객 데이터를 활용해 마케팅 활동을 활성화하는 고객관계관리(CRM) 활동을 1990년대부터 시작했다.

CRM은 기업이 보유하고 있는 데이터를 통합하는 데이터웨어하우스(DW), 고객 데이터 분석(Data Mining)을 통한 고객유지와 이탈방지 등과 같은 다양한 마케팅 활동을 진행하는 것을 뜻한다.

이러한 고객분석은 빅데이터 시대를 맞이해 전환점을 맞고 있다. 분산처리방식과 같은 빅데이터 기술을 활용해서 과거와 비교가 안 될 정도의 대규모 고객정보를 빠른 시간 안에 분석하는 것이 가능하다. 트위터와 인터넷에 생성되는 기업 관련 검색어와 댓글을 분석해 자사의 제품과 서비스에 대한 고객 반응을 실시간으로 파악해 즉각적인 대처를 시행하고 있다.

소프트웨어나 하드웨어도 오픈 소스 형태의 하둡(Hadoop)이나 분석용 패키지인 R과 분산병렬처리기술, 클라우드 컴퓨팅 등을 활용하면 기존의 비싼 스토리지와 데이터베이스에 기반한 고비용의 데이터웨어하우스를 구축하지 않더라도 효율적인 시스템 운용이 가능하다[7].

국내 DB산업 규모는 현재 11조 6000억원에 달한다. 이중 DB분석 솔루션 시장은 연평균 10.2%의 성장률을 기록하면서 2013년 국내 DB 산업 규모는 2조원을 돌파한 것으로 나타났다[8].

DB컨설팅 및 솔루션 시장은 공공에 비해 민간부문의 사업 비중이 매우 큰 영역으로, 신속하고 정확한 데이터 처리가 기업생존과 직결되는 금융권과 제조업의 도입이 활발하기 때문에 그 성장세가 지속되고 있는 것으로 분석하고 있다[9]. 그런데, 시장에서는 외국계 기업이 약 80%를 점유하며 강세를 보이고 있다. 특정 외국계 벤더의 DBMS 점유율이 60%에 이르고 있으며, 공공 시장의 경우 시장 점유율이 90%에 육박하는 상황 속에서도 국내 업체들은 가격 및 커스터마이징, 유지보수 경쟁력을 기반으로 시장에서의 경쟁력을 높여 왔다. 그러나, DBMS가 함께 제공되는 데이터웨어하우스(DW) 어플라이언스 제품은 기존 데이터웨어하우스(DW)보다 가격도 저렴하고 성능 및 유지보수 편리성이 뛰어나기 때문에 국내 DBMS 업체들의 시장입지를 위축시킬 수 있다.

Ⅳ. 결 론

Gartner에서는 2013년도 10대 전략기술에서 전략적 빅데이터를 꼽았다. '12년의 빅데이터 애기가 빠지지 않았다. 다른 점이 있다면 단순히 소셜 데이터를 긁어모아 고객 데이터와 결합해 결과를 바라보기보다는 하둡과 같은 NoSQL을 통해 비정형 데이터 그 자체를 분석하는 게 중요해졌다. 가트너는 또한 미래에는 하나의 DW가 기업의 모든 의사 결정을 지원할 수 있는 통합된 시스템으로 진화할 것이라고 전망했다.

본 논문에서는 기존의 DW어플라이언스 제품 및 기술 동향에 대해 알아보고 또한 기존 빅데이터 분석기술을 분석하여 향후 빅데이터 분석처리 기술에 대한 방향을 연구하였다.

결론적으로 DW를 구축하던 기존 방식과 달리 DW어플라이언스는 서버, 스토리지, DBMS를 각각 도입하여 튜닝 작업 감소 및 성능 향상, 도입비용 절감, 데이터 모델링의 축소, DBMS 및 OS의 업그레이드 비용 절감, 데이터베이스 아키텍처 관리의 편리성이 증대의 장점을 갖는다.

이것은 기존 데이터웨어하우스(DW)보다 가격이 저렴하고 성능 및 유지보수 편리성이 뛰어나다. 하지만 새로운 데이터 분석 기술을 결합한 어플라이언스 또한 개발할 필요가 있다.

따라서, 국가적인 R&D 지원사업을 통하여 글로벌 기업들과 거뜬히 경쟁할 수 있는 국산 DBMS와 서버-스토리지 플랫폼 기반의 새로운 DW 어플라이언스 개발을 조속히 진행해야 할 것으로 판단된다.

사사의 글

본 논문은 2013년도 지식경제부의 SW 전문인력양성 사업의 재원으로 정보통신산업진흥원의 고용계약형 SW석사과정 지원 사업(HB301-13-1003)으로부터 지원받아 수행되었습니다.

참고문헌

[1] 김정태, 빅데이터 핵심 기술 및 표준화 동향, 정보통신 동향분석 제28권 제1호, 2013년 1월

[2] 조성우, Big Data 시대의 기술, 중앙연구소, 2012년

[3] 정용완, DW 어플라이언스 기술 동향 분석, 한국정보기술학회지 제10권 제2호, 2012년

[4] 정윤철, ETL상에서 파일 시스템을 이용한 대용량데이터 처리 기법, 고려대학교 컴퓨터정보통신대학원, 2009년

[5] www.oracle.com, Oracle

[6] www.sap.com, SAP

[7] 정용찬, 빅데이터, 커뮤니케이션북스, 2012년

[8] 한국데이터베이스진흥원, "2013년도 국내 데이터베이스 산업 시장분석 결과보고서", 2013년 1월.

[9] 정보통신산업진흥원 SW 정책팀, "2012년 SW 10대 비즈니스 및 기술이슈 요약". 2012년 1월.