빅데이터 시스템의 데이터 수집 및 저장에 관한 연구

박지훈, 김경환, 정은수
부경대학교 컴퓨터공학과
ghunjjang@naver.com, sharp_line@naver.com, vvwvvu@gmail.com

A Study on the Data Collection and Storage of Big Data Systems

Jihun-Park, Gyunghwan-Kim, Eunsu-Jung
Dept of Computer Engineering, Pukyong national University

요 약

빅데이터는 저장되지 않았거나 저장되더라도 분석되지 못하고 버리게 되는 방대한 양의 데이터를 말한다. 실제로도 빅데이터는 페이스북, 트위터등의 소셜 네트워크에서 많이 발생하고 있는데, 이러한 방대한 데이터들을 어떻게 효율적으로 저장하고 분석하는지에 대한 관심이 많아지고 있다. 따라서 본 논문에서는 빅데이터의 개념, 빅데이터의 향후 동향과 이슈들에 대해 살펴보고, 빅데이터 시스템이 데이터를 수집하고 저장하는 것에 대한 고려할만한 사항들과 효율적인 해결방안에 대해 제시하였다.

<keyword> 빅데이터, 빅데이터 저장, 데이터 수집, 클라우드 시스템, 데이터 거버넌스, 데이터 아키텍쳐

Ⅰ. 서론

빅데이터 분석은 몇 년 전 클라우드 컴퓨팅 등과 함께 큰 주목을 받으며 등장한 이후, 이제는 데이터를 활용한 분석을 대신하는 통상적인 표현으로 자리잡았다. 그러나 빅데이터 분석이라는 표현이 빈번하게 사용되고 큰 기대를 받았지만, 최근 실시된 NIA (한국정보화진흥원)의 조사결과 등에 따르면 아직까지 국내에서는 수집 및 저장 뿐만 아니라, 활용 수준이나 기반, 인프라와 제도 정비 등 수준에서 기대에 미치지 못 하고 있는 상황이다. 빅데이터는 저장되지 않았거나 저장되더라도 분석되지 못하고 버리게 되는 방대한 양의 데이터를 말한다. 관리해야 할 데이터가 늘어나면 그만큼 저장장치 용량 증설도 마찬가지로 필요하다. 대규모 데이터를 효율적으로 관리 및 저장하기 위해서 빅데이터 저장 기술이 연구되고 있으며, 그 필요성이 큰 것으로 판단된다. 이에 본 고에서는 국내외의 빅데이터 필요성과 저장 및 수집의 측면에서의 효율적인 해결방안을 짚어보고자 한다.

Ⅱ. 빅데이터 국내외 시장동향과 기술이슈

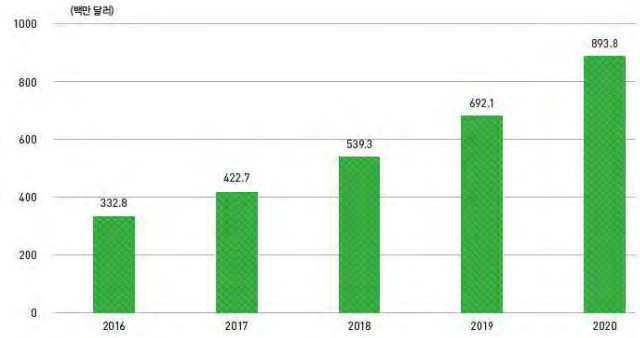
1. 세계 시장 동향

세계 빅데이터 시장은 현재도 성장 중이며, 시장조사기관 마다 규모의 차이는 다소 있으나 공통적으로 높은 성장률을 예측한다. IDC의 자료에 의하면 빅데이터 시장을 크게 인프라, 소프트웨어, 서비스 등 3가지로 분류하고 모두 성장할 것으로 전망하며, 전 세계 빅데이터 인프라 시장은 2019년 까지 486억 달러 규모(연평균 성장률 23.1%)에 이를 것으로 전망된다. 미국 지디넷은 빅데이터 및 분석 시장이 2019년 까지 1,879억 달러 규모(연평균 성장률 50%)로 성장할 것으로 전망된다.

2. 국내 시장 동향

국내 빅데이터 산업은 아직까지는 도입 초기 수준이나 기업 전반에서 실질적인 인프라를 구현하려는 단계로 접어들었다. 2015년 기준 국내 빅데이터 시장규모는 2014년 대비 30% 성장한 2,623억 원 규모이며, 한국과학기술정보 연구원에 의하면 국내 빅데이터 시장은 2020년 까지 8억 9천만 달러(한화 약 1조 원)규모까지 성장할 것으로 예측된다. 빅데이터 관련 정부투자 또한 2013년 230억원에서 2015년 기준 698억 원으로 세 배 이상 증가했다.

<그림 1> 세계 빅데이터 시장 동향 및 전망
출처 : Statista, Forecast of Big Data market size, based on revenue, from 2011to 2016, 2016 Wikibon, Big Data Market Forecast, 2011-2026



<그림 2> 국내 빅데이터 시장 전망
출처 : 한국과학기술정보연구원(KISTI)

3. 수집에 대한 이슈

업무 처리과정에서 자연적으로 발생하는 부산물 데이터 또는 사람들이 접근하는 휴리스틱한 데이터들은 계획에 따라 설계해서 수집한 데이터가 아니다. 페이스북, 트위터 등 소셜 네트워킹 서비스의 확산으로 데이터베이스에 잘 정리되어 저장되는 데이터가 아닌 웹문서, 메일, 소셜 데이터와 같은 비정형 데이터이다. 이런 비정형 데이터들은 저장공간에 실시간으로 수집되어 저장이 된다.

알고리즘만 잘 짜여져 있으면 데이터 수집은 무한정 가능하다. 하지만 데이터 생산량의 증가를 감당하지 못한다. 수집에도 한계가 존재하며, 방대한 데이터들 중에서도 유용한 데이터가 부족해 데이터 부재와 품질저하에 관한 문제점도 야기된다. 그러므로 분석과 해석을 하고 적용을 하는 어려움이 있다. 비정형 데이터들을 가공해 객관성과 체계성, 부하량 등을 개선시킬 필요성이 있다. 또한 대표성을 가진 수집이 필요하다.

4. MapReduce 프로그래밍 모델

MapReduce는 프로그래머로부터 시스템 수준의 세부 사항을 숨길 수 있는 추상화를 제공한다. 이는 독립적인 작업으로 대량의 데이터 집합을 병렬 및 분산처리를 할 수 있도록 한다.

MapReduce 프로그래밍 모델은 Lisp 및 ML과 같은 함수 프로그래밍에 기반을 두고 있다. key-value의 쌍은 MapReduce의 기본 자료구조를 구성하며, key와 value는 정수, 실수, 문자열, 바이트열 또는 임의의 복잡한 자료구조로서 정의될 수 있다. MapReduce 프로그램은 Map과 Reduce의 함수로 이루어진다.

Map 함수는 주어진 블록의 데이터를 읽고 (key1, value1) 응용 처리를 수행한 후 기본 자료구조인 key-value의 쌍인 (key2, value2)의 리스트를 생성한다. Reduce 함수는 동일한 중간 key와 관련된 모든 값을 key-value의 쌍인(key2, list(value2))으로 입력받아 새로운 key-value 쌍인 (key3, value3)의 리스트를 출력한다. 이 과정에서는 중간 값에 대한 shuffle과 sort를 통한 그룹핑을 수행한다.

Ⅲ. 빅데이터 해결 방안

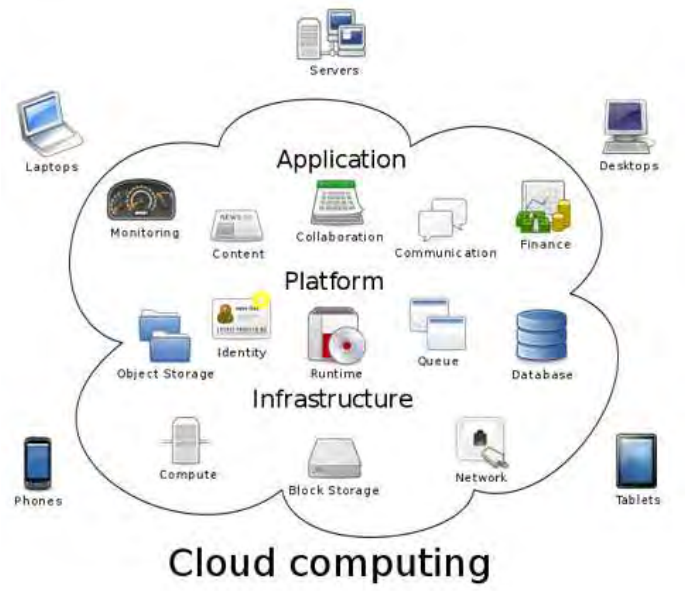
향후 사물 인터넷이 확산되고 개인 데이터가 증가하면서 데이터 수집과 실시간 서비스 제공으로 다변화 할 것으로 예상된다. 증가하는 데이터의 처리 및 활용을 위해서는 다양한 유형의 데이터의 효율적 관리, 빠른 속도, 정확한 처리가 필요하다.

페타바이트(peta byte) 수준의 데이터를 효과적인 관리를 위해 클라우드 컴퓨팅을 이용한 빅데이터 분산 처리 구조를 제안한다.

대용량 데이터를 중앙 집중식으로 관리하기보다는 기관 별로 클라우드 기술을 이용하여 다양한 위치에 클라우드 센터들을 구축하고 구축된 데이터 센터들을 하나의 데이터 센터처럼 관리할 수 있도록 네트워크 기술인 SDN을 통해 지원하여, 시간과 위치에 제약 받지 않고 데이터 처리와 분석에 필요한 컴퓨팅 자원과 데이터 저장 스토리지에 접근 할 수 있게 하는 구조 실시간으로 처리해줄 수 있는 장점이 있다.

또한 일반적인 클라우드 컴퓨팅인 IaaS 방법외에 개발자가 개발에 필요한 서버, 저장소, 네트워크 및 데이터베이스의 기본 인프라를 설정하여 관리할 필요 없이 더 쉽고 빠르게 생산할 수 있도록 디자인 되어있다.

다양한 데이터를 공유 시 데이터 복제 및 중복을 최소화하면서도 실시간으로 원하는 정보를 찾아 그에 맞는 서비스를 제공할 수 있는 특징을 가진다. 이는 각 물리 서버에 접근하기 위한 각각의 접근 정책 및 제어 기법이 필요하지 않고 통합 클라우드 관리 플랫폼에 의해 관리할 수 있어 보안 및 접근 관리가 편리한 이점을 가진다.



<그림 3> 클라우드 컴퓨팅
출처 : commons.wikimedia.org

클라우드 컴퓨팅을 이용하면 공용 클라우드와 사설 클라우드 간에 데이터와 응용 프로그램을 공유할 수 있는데, 그 기술은 함께 바인딩된 공용 클라우드와 사설 클라우드를 결합한 하이브리드 클라우드라 한다. 하이브리드 클라우드에서는 사설 클라우드와 공용 클라우드 간에 데이터와 응용 프로그램이 이동할 수 있도록 하여 기업에 더 큰 유연성과 다양한 배포 옵션을 제공한다.

사업자 및 개발자들도 단일의 플랫폼에서 서비스를 개발하고 정보를 수집할 수 있기 때문에 비즈니스 목표를 달성하기 위한 시간 단축 및 다양한 형태의 신사업이 도출될 것이며, 비용과 확장성 측면에서도 유리하다. 현재 우리나라에서도 고객 요구 분석과 개발 과정 관리 기업이 활용하고 있으며 정부 주도로 공공 데이터를 공개하고 점차 데이터 공유 및 활용에 대한 요구가 증가하고 있으므로 클라우드 컴퓨팅이 정보 공유의 장이 될 것으로 보인다.

다음으로, 데이터의 빠른 처리 성능 및 정확성을 위한 방안으로 데이터 인덱싱 기술을 제안한다.

데이터를 효율적으로 조회할 수 있도록 도와주는 데이터 구조인 인덱스는 성능 향상에 있어 필수적이라고 할 수 있는데, 데이터의 양이 커질수록 중요해진다. 데이터의 크기가 작고 부하가 적은 데이터베이스에는 인덱스가 없더라도 작동이 되지만, 데이터의 집합이 커질수록 인덱스가 없다면 성능이 크게 떨어지게 되므로 빅데이터에서 인덱스는 반드시 필요하다.

인덱스는 서버 계층에서 구현이 되지 않고, 스토리지 엔진 계층에서 구현되므로 표준화가 되어 있지 않다. 엔진마다 조금씩 다르게 작동하여, 인덱스의 특징을 알고 사용목적에 알맞은 인덱스를 사용해야 한다.

빅데이터 처리 프레임워크인 Hadoop에서도 인덱스를 이용한 연구가 진행되고 있는데, 다차원의 인덱스를 Hadoop에서 구현하여 그 효과를 성능평가를 통해 입증하였다.

빅데이터 인덱싱 관련기술로는 대표적으로 루씬과 솔라가 있다. 둘은 확장 가능한 고성능 정보검색 라이브러리로서, 문서를 색인하고 색인된 문서를 검색하는 기능을 제공한다. 다중 데이터들의 획득과 병합 및 구문분석과 더불어 데이터의 인덱싱 및 검색을 위한 준비 후에 사용자와의 상호작용을 수행하여 질의 검색에 따른 인덱싱 및 검색기능을 제공한다. 사용자 질의에 대한 결과를 분석단계로서 질의 결과는 사실대로 표현되며, 필요시 다양한 형태의 그래프들을 제공하여 분석에 도움을 줄 수 있다.

또한 빅데이터를 효과적으로 활용하기 위한 메타데이터의 활용분야들의 지속적인 활성화를 위해서 메타데이터 관리체계의 표준화가 선행되어야한다 표준화 작업은 외부 조직 간의 정보 호환성과 공유를 용이하게 하여 단체 간의 국제 교류를 향상 시킬 수 있다.

현재 다양한 분야의 표준화 위원회에서 개별적인 표준화를 위해 노력하고 있는데 이들은 표준의 대상 성격 목적에 따라 데이터 요소 관련 표준 영역중심 표준 네트워크 자원기술 메타데이터로 분류 할 수 있다.

빅데이터의 메타데이터 관리 프레임워크를 위하여 이와 연관되어 빅데이터의 메타데이터를 정의하고 데이터를 구조화 하는 빅데이터 정의 및 구축 기술, 메타데이터를 관리하고 공유하기 위한 메타데이터

레지스트리 기술 이러한 기술들을 기반으로 빅데이터를 수집할 수 있는 영역을 특정하고 빅데이터를 위한 메타데이터관리 프레임워크의 표준화를 제안한다.

표준화 방안은 급속히 확산되고 있는 IoT, SNS 등 다양한 빅데이터 생성 객체들을 효과적으로 통합할 수 있는 응용 서비스 개발에 활용될 수 있을 것으로 보이며, 빅데이터 활용 및 응용 개발을 위한 규범적인 메타 데이터 관리 인프라 제공 빅데이터 관리의 체계성 및 정확성 보장하는 효과가 기대된다.

Ⅳ. 결론 및 향후연구

본 고에서는 빅데이터 수집과 저장에 관한 문제와 응용에서의 빅데이터를 처리하기 위한 연구를 진행하였다. 빅데이터를 소개하고 활용을 위한 기술인 MapReduce programming model에 대해 설명하였다. 한국과학기술정보연구원(KISTI)를 통해 2016년 이후 수집되어진 빅데이터의 동향을 통해 빅데이터 전망을 확인할 수 있었으며, 최근의 인공지능과 사물인터넷과의 결합을 통해 그 중요성이 크다는 것을 알 수 있었다. 그에 따라 네트워크 품질 향상에 대한 요구가 지속적으로 증가할 것이므로 고품질 광대역의 통합망이 반드시 필요하다는 시사점을 제시하였으며, 그 해결책으로 효율적으로 수집하고 저장할 수 있는 방법으로 데이터 센터 기반의 클라우드 컴퓨팅 기술을 제시하였다.

현재 4차 산업형명 시대의 도래를 앞둔 상황에서 빅데이터 수집과 저장뿐만 아니라 빅데이터 분석 활용수준은 전반적으로 재고되어야 한다. 앞으로 바뀌고 있는 빅데이터 분석의 명확한 활용 목적 수립과 활용 가능한 데이터에 대한 정확한 파악 및 추가 데이터 확보등에 노력을 기울일 때 인덱싱을 통한 빅데이터 분석 활용이 활성화 될 것으로 보고 연구를 진행할 계획이다.

Ⅴ. 참고문헌

[1] Guo, Z.(G.), R. Singh, and M. Pierce, "Building the PolarGrid Portal Using Web 2.0 and OpenSocial" The International Conference for High Performance Computing, Networking, Storage and Analysis (SC'09), Porland, OR, ACM Press, pp. 5, 11/20/2009.

[2] 정우진 "빅데이터를 말하다" 클라우드북스 2013

[3] 하정욱, " 클라우드와 빅 데이터", IDG Summary, http://www.itworld.co.kr/techlibrary, 2012.06.28.

[4] Wikipedia, http://en.wikipedia.org/wiki/Big_data, 2012.06.28.

[5] Yunhee Kang, Heeyeoul Choi, An Empirical Study for Handling Scientific Datasets, International Journal of Grid and Distributed computing, Vol 5, No 3, 2012

[6] J. Ekanayake, et al., "MapReduce for Data Intensive Scientific Analyses," the 2008 Fourth IEEE International Conference on eScience 2008.

[7] 정우진, 글로벌 혁신 기업의 Digital Transformation, 2016

[8] Digieco, 빅데이터의 이해와 활용, 2015

[9] Haojun Liao., Jizhong Han., Jinyun Fang. "Multi-dimensional Index on Hadoop Distributed File System", 2010 Fifth IEEE international Conference on Networking, Architecture, and Storage, 2010

"본 논문은 2017년 한이음 ICT멘토링 프로젝트의 결과물입니다."