빅 데이터 분석 플랫폼 서비스 동향

박변용*, 김성수**, 강정호***, 전문석*
*숭실대학교 컴퓨터학과
**한국정보화진흥원
***배화여자대학교 정보보호과
e-mail:*qusdyd0905@naver.com **cryptoauth@nia.or.kr
***kangsamm@baewha.ac.kr *mjun@ssu.ac.kr

Trend of Big data Analysis Platform Service

Byeon-Yong Park*, Sung-Soo Kim**, ***Jeong-ho Kang, Moon-Seog Jun*
*Dept. of Computer Science, SoongSil University
**National Information Society Agency
***Dept. of Information Protection, BaeWha Women's University

요 약

인터넷이 발달함에 따라 데이터의 생산량은 기하급수적으로 증가하고 있고, 생성된 막대한 양의 데이터를 사용하는 목적에 맞게 분석하여 이익이 될 수 있는 유의미한 정보를 얻을 수 있다. 본 논문에서는 빅 데이터 분석을 위한 여러 가지 기술들과 분석 플랫폼 동향을 알아보고, 국내에서 빅 데이터가 발전하기 위한 방안에 대해서 알아본다.

1. 서론

빅 데이터(Big data)란 통상적으로 사용되는 데이터 수집, 관리 및 처리 소프트웨어의 수용 한계를 넘어서는 크기의 데이터를 의미한다. 빅 데이터는 3V를 만족하여야 하는데 데이터의 규모(Volume), 데이터의 다양화(Variety), 데이터를 빠른 속도로 처리(Velocity) 의 3가지를 만족해야 한다. 규모란 데이터의 크기를 나타내는 것이고, 다양화란 수집되는 데이터가 정형, 비정형의 다양한 형태로 수집될 수 있는 것을 의미하는 것이다. 빅 데이터가 등장하게 된 이유로는 정보화 시대를 맞이하여 SNS(Social Network Service)의 발전과 함께 데이터의 생산량은 기하급수적으로 증가하게 되었다. 시장조사기관 IDC에 따르면 2011년에는 1.8 제타바이트, 2012년에는 2.7 제타바이트의 데이터가 생산된 것과 같이 데이터의 생산량은 매년 증가하고 있으며 2025에 생성되는 데이터의 양은 약 170 제타바이트만큼 증가할 것이라고 예상하고 있다[1].

(그림 1) 전세계 데이터 생산량 추이(IDC, 2017)

(그림 2) 전 세계 SNS 사용자 수(Statistica, 2017)

그림 1은 미국의 시장 조사 회사인 Gartner에서 전세계 데이터 생산량을 보여주고 있고, 그림 2는 Statistica에서 발표한 전 세계 SNS 사용자 수를 보여주고 있다. 데이터의 생산량이 기존과는 비교할 수 없게 많아져 기존의 데이터베이스로는 처리하기가 힘들어졌기 때문에 새로운 빅 데이터 분석 플랫폼을 필요로 한다. 여기서 중요한 것은 전 세계에서 생성되는 많은 데이터들을 분석하여 유의미한 정보를 얻을 수 있고 기업에서는 분석된 데이터로 의사결정을 하는 등 많은 이익을 창출해낼 수 있다. 따라서 본 논문에서는 빅 데이터 분석 플랫폼에 대해 살펴본다.

2. 관련 기술

빅 데이터 분석 플랫폼은 데이터를 목적에 맞게 분석할 수 있는 여러 가지 분석기술들이 있다. 크게 엑셀 파일과 같은 구조화된 데이터를 분석하는 정형 데이터 분석과 사진, 동영상, SNS 등의 정형화 되지 않은 데이터를 분석하는 비정형 데이터 분석이 있다. 본 장에서는 분석 기술의

종류에 대해 살펴본다.

데이터 마이닝 : 대규모의 데이터를 데이터 간의 특정 패턴, 관계 등을 추출하여 가치 있는 정보를 찾는 기술이다. 정형 데이터를 다루고 기계학습, 패턴인식, 통계학 등 기술들을 이용한다.

텍스트 마이닝 : 비정형 텍스트를 정형화시키고 자연어 처리 기술을 기반으로 하는 특정 단어와 문맥의 연관성을 분석하여 가치 있는 정보를 찾는 방법이다.

오피니언 마이닝 : 텍스트 마이닝을 기반으로 하고 여론을 긍정 ,부정, 중립으로 구분하여 분석하는 기술이다. 텍스트 마이닝과는 다르게 감성분석기법을 이용하여 감정을 파악한다.

SNS 분석 : 네트워크의 연결 구조 및 연결 강도 분석을 바탕으로 네트워크상 영향 요소를 찾는데 활용되는 방식이다.

다음은 McKinsey에서 발표한 빅 데이터 분석기법이다[2].

<표 1> 빅 데이터 분석기법

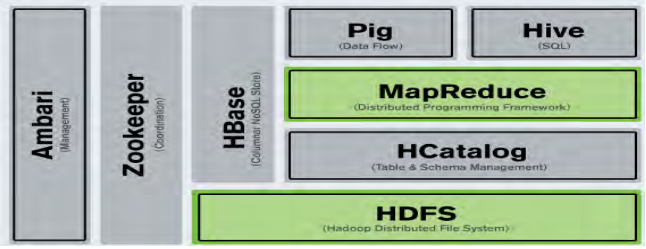
빅 데이터 분석기법	• A/B testing(A/B 테스팅) • Association rule learning(연관규칙학습) • Classification(분류) • Cluster analysis(클러스터 분석) • Data fusion and data integration(데이터 퓨전/데이터 통합) • Data mining(데이터 마이닝) • Ensemble learning(앙상블 학습) • Genetic algorithms(유전자 알고리즘) • Machine learning(기계학습) • Natural language processing(자연어처리) • Neural networks(신경망) • Network analysis(네트워크 분석) • Optimization(최적화) • Pattern recognition(패턴 인식) • Predictive modeling(예측 모델링) • Regression(회귀분석) • Sentiment analysis(감성 분석) • Signal processing(신호 처리) • Spatial analysis(공간 분석) • Statistics(통계) • Supervised learning(지도 학습) • Simulation(시뮬레이션) • Time series analysis(시계열 분석) • Unsupervised learning(비지도 학습) • Visualization(시각화)

3. 빅 데이터 분석 플랫폼 및 동향

빅 데이터를 위한 분석플랫폼은 데이터를 수집, 저장, 처리, 분석, 가시화하는 것뿐만 아니라 고성능의 컴퓨터 인프라가 필요하게 된다. 빅 데이터를 수행하기 위한 고성능의 인프라에는 분산 컴퓨팅 기술, 고성능 컴퓨터 기술, 인메모리 컴퓨팅 기술 등의 컴퓨팅 인프라가 필요하다.

3-1 오픈소스 플랫폼 서비스

본 장에서는 분산처리를 위한 기술인 하둡(hadoop)에 대해 설명한다. 하둡은 아파치(apache)에서 만든 오픈소스 플랫폼이며 빅 데이터 환경에서의 분산 저장, 처리를 위한 시스템이다[3]. 하둡은 수집된 데이터를 여러 대의 서버에 분산 저장하는 HDFS(Hadoop Distributed File System)와 분산 저장된 데이터를 병렬로 처리하는 MapReduce로 구성된다. 하둡에서 이 2가지 기술이 핵심이고 그림 3과 같다. 분산 컬럼기반의 데이터베이스이고 HDFS의 데이터에 대한 실시간 read/write 기능을 제공하는 것인 HBase, 하둡 데이터를 SQL ,HiveQL을 이용하여 다룰 수 있게 해주는 기술인 Hive, 복잡한 MapReduce 프로그래밍을 대체할 수 있는 Pig, 분산 환경에서 서버들 간 상호 조정에 대한 서비스를 제공하기 위해 동기화, 통합 관리를 하는 ZooKeeper와 같은 요소들로 구성된다.



(그림 3) 하둡의 구성요소

3-2 상용 플랫폼 서비스

상용 플랫폼 서비스는 비즈니스용 기능 집합을 제공하는 것이며, 빅 데이터에서의 상용 분석처리 플랫폼은 대부분 오픈소스인 하둡을 기반으로 하여 플랫폼을 만들었다.

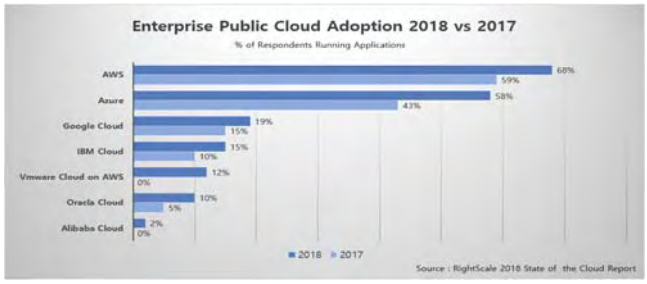
최근 빅 데이터에서 상용 분산처리 플랫폼으로 사용되는 것으로는 IBM사의 Big data플랫폼, Amazon의 플랫폼인 AWS, Microsoft의 플랫폼인 Azure 등이 하둡 기반의 클라우드 서비스를 사용하는 플랫폼들이다. IBM에서는 빅 데이터 분산 처리를 위해서 BigInsight를 제공하는데 여기서 Big R, Text Analysis, Big SQL, BigSheet라는 도구로 데이터의 분산처리를 할 수 있다. AWS에서 분산처리를 위한 서비스로는 Amazon Elastic Map Reduce, Elasticsearch Service, Athena를 가진다. Azure에서는 빅 데이터 분산 처리를 위한 서비스로 HDInsight, Data Lake Analytics, Machine Leaming studio를 사용한다.

<표 2> 상용 플랫폼 분석도구

플랫폼	분석을 위한 도구
AWS	Amazon Elastic Map Reduce Elastic search Service, Athena
Azure	HDInsight, Data Lake Analytics Machine Learning studio
IBM	Big R, Text Analysis Big SQL, BigSheet

3-3 분석플랫폼 동향

상용 플랫폼은 각 회사 플랫폼에서 하둡의 분산처리를 더 효율적으로 하기 위해 클라우드 컴퓨팅 환경을 제공할 수가 있다. 클라우드란 데이터를 저장할 때 컴퓨터 내부가 아닌 인터넷 상의 서버에 영구적으로 저장이 되는 공간을 의미한다, 클라우드 컴퓨팅은 이런 인터넷 상의 서버에 저장된 데이터를 각종 IT 기기를 이용하여 언제 어디서든지 이용할 수 있게 하는 기술의 의미한다. 클라우드를 사용할 경우 여러 가지 장점이 생기기 때문에 빅 데이터를 플랫폼을 개발하는 회사에서는 서비스에 필요한 기능과 클라우드 컴퓨팅을 사용하여 빅 데이터 분석 플랫폼을 구축할 수 있다.



(그림 4) 기업 클라우드의 채택(RightScale,2018)

그림 4의 설문조사는 RightScale에서 클라우드 컴퓨팅 채택에 대한 광범위한 조직의 기술 전문가들에게 기업에서의 빅 데이터 플랫폼사용에 관한 것을 나타낸다. Alibaba와 같은 새로운 플랫폼의 등장도 확인할 수가 있다. 그림으로 알 수 있는 것은 최근 빅 데이터 플랫폼은 이용률로 보았을 때 AWS, Azure가 기업에서 가장 많이 사용되고 있고 나머지 플랫폼들이 추격하고 있는 것이라고 볼 수 있다.AWS, Azure이 강제를 보이는 이유에 대해서 AWS는 높은 확장성, 대규모 서비스 지원, 솔루션 제공 등의 장점이 있다. Azure은 보통 기업들이 Window환경을 사용하기 때문에 소프트웨어 호환성이 좋아 사용을 하고 하이브리드 클라우드, 직관적인 UI등의 장점을 가지기 때문이다. 이 플랫폼들은 IaaS(Infra as a Service)인데 IaaS란 클라우드 컴퓨팅의 서비스 유형을 나타내는 것으로 서버, 저장장치, 네트워크 장비와 같은 기술만 제공하는 것을 인프라 서비스(IaaS)라고 한다. 이 외에도 인프라와 운영체제, 플랫폼 개발을 위한 도구까지 제공하는 것을 플랫폼 서비스(Platform as a Service), 완성된 애플리케이션이나 응용프로그램까지 클라우드로 제공되는 소프트웨어 서비스(Software as a Service) 3가지 유형으로 나뉜다. 그 외에 플랫폼은 SAP 의 HANA Cloud 플랫폼, 시스코의 CPA와 같은 다른 플랫폼들이 있다.

4. 결론

현재까지 생성된 전체 인터넷의 데이터의 90% 이상이 최근 2년 사이에 생성이 되었고 미래에 AI, IoT와 같은 기술이 발달함에 따라 데이터의 생산량이 2025년에는 약 170 제타바이트만큼 생산이 될 것이라 예상이 되고 있다. 이러한 이유로 빅 데이터 분석 플랫폼의 중요성은 데이터의 생산량이 증가함에 따라 더욱 커질 것이다. 빅 데이터는 이미 기업에서는 의료시스템, 교통시스템, 문화산업 등의 여러 방면에서 사용되고 있으며 온라인 쇼핑몰에서 고객의 감성을 분석하여 원하는 정보를 알려주고, 새벽시간 서울 시내에서 발생한 통신 데이터들을 분석하여 심야버스를 운행하는 사례에서 볼 수 있듯이 빅 데이터의 활용성은 높다. 그렇지만 현재 우리나라에선 빅 데이터에 관한 전문 인력의 수가 부족하기 때문에 세계적으로 경쟁력을 높이기 위해서 전문 인력을 체계적으로 양성하는 방안이 마련되어야 한다.

참고문헌

[1] 씨게이트 “Data Age 2025” www.DataAge2025.com
[2] McKinsey Global Institute, <Big data : The next frontier for innovation, competition, and productivity>
[3] Apache Hadoop "Definition of Hadoop" http://hadoop.apache.org
[4] 김도균, 최진영 “실시간 빅 데이터 분석을 위한 플랫폼 제공서비스 동향”. ie 매거진, 23(4),23-27
[5] 김재생 “빅 데이터 분석 기술과 활용사례” 한국콘텐츠학회지, 12(1), 14-20
[6] 안춘모 “빅 데이터 플랫폼 현황 및 이슈 분석” insight report 2017-33