



미래사회와 빅 데이터(Big data) 기술

정병권

ETRI 서버플랫폼연구팀/선임연구원

bkjung@etri.re.kr

김학영, 최완

ETRI 클라우드컴퓨팅연구부

1. 서론
2. 빅 데이터 요소기술
3. 빅 데이터 분석기술
4. 빅 데이터 처리기술
5. 빅 데이터 미래기술
6. 결론

1. 서론

스마트폰과 SNS 혁명으로 인해 몇 년 전만해도 생각지도 못한 엄청난 양의 데이터가 생성되고 있다. 양만 많은 것이 아니라 그 종류도 다양해지고 있으며, 정형화된 데이터뿐만 아니라 비정형화된 데이터도 늘고 있다. 페타 바이트급 데이터 웨어하우스, 소셜 네트워크, 실시간 센서 데이터, 지리 정보 및 기타 여러 가지 새로운 데이터 소스가 출현함에 따라 기업들은 다양한 문제에 직면하게 되었다. 이러한 데이터의 급격한 증가는 이제 기존 처리방식으로는 증가하는 데이터를 감당할 수 없으며, 정보처리의 새로운 패러다임을 필요로 한다. 빅 데이터(Big Data)란 무엇인가? 일반적으로 빅 데이터는 기존 데이터에 비해 너무 커서 기존 방법이나 도구로 수집, 저장, 검색, 분석, 시각화 등이 어려운 정형 또는 비정형 데이터를 의미한다. 두 개의 기관에서는 다음과 같이 정의하였다[1].

- DB의 규모에 초점을 맞춘 정의(Mckinsy, 2011): 일반적인 DB SW가 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터
- 업무수행에 초점을 맞춘 정의(IDC, 2011): 빅 데이터는 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고, 데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처

* 본 내용과 관련된 사항은 ETRI 서버플랫폼연구팀 정병권 선임연구원(042-860-1537)에게 문의하시기 바랍니다.

** 본 내용은 필자의 주관적인 의견이며 NIPA의 공식적인 입장이 아님을 밝힙니다.

빅 데이터의 정의를 단순한 정량적인 차원에서 접근해서는 안될 것이다. 왜냐하면 빅 데이터는 지속적으로 변하면서 산업별, 시장별 구분에 따라 다르게 적용되기 때문이다. 2011년에 발표된 IDC의 Digital Universe 연구조사에 의하면, 새롭게 생성되거나 복제된 정보의 양이 2011년에 1.8 ZB(1조 8,000억 Gigabyte)를 넘어서고, 향후 5년 후에는 거의 9배 가까이 증가할 전망이다. 1.8 제타 바이트는 대한민국 모든 사람(약 4,875만 명, 2010년 기준)이 18만 년 동안 쉬지 않고, 1분마다 트위터에 3개의 글을 게시하는 양과 같다. 또 2시간짜리 HD 영화 2,000억 개와 맞먹는 양이다. 전세계의 디지털 정보량은 2년마다 2배씩 증가하고 있다[2].

소셜미디어와 스마트폰에서 촉발된 빅 데이터 이슈가 산업 전반에 확산되면서 방대한 양의 데이터가 생산되고 있고, 이를 기반으로 한 데이터 중심의 제4세대 연구 패러다임이 새롭게 떠오르고 있다. 기업의 무한 경쟁, 과학기술의 거대화화 및 융복합화 가속, 삶의 질적 수준 향상 요구, 재난 대응, 일자리 창출 등 사회 변화와 현안 해결에 빅 데이터의 역할이 매우 중요하다. 또한 빅 데이터의 수집과 분석을 위한 컴퓨팅 파워, 소프트웨어, 정보 분석 모델, 가시화, 인공지능 등의 기술이 수반되어야 할 것이다.

2. 빅 데이터 요소기술

빅 데이터를 설명할 때 다음과 같이 크게 3가지 요소를 들 수 있다[3].

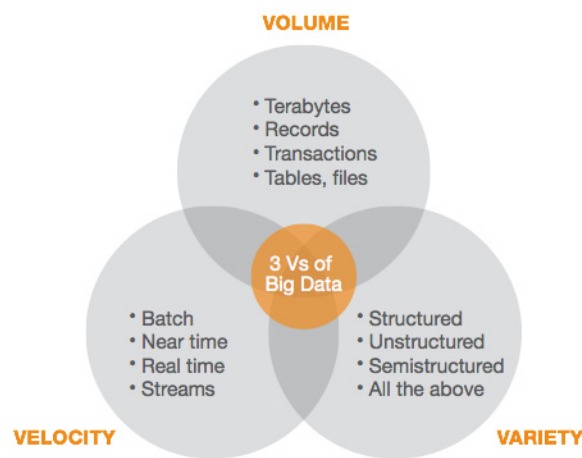
첫째, 데이터의 크기(Volume): 데이터의 크기로 물리적인 크기 보다는 앞서 정의에서 설명한 크기이다. 웹 로그 데이터나 g-mail 등의 메일 데이터는 수 PB 이상이 되지만 트위터 네트워크 데이터는 수십 GB 미만이다. 앞의 데이터는 안정적인 저장이 가장 큰 해결 과제인 반면 네트워크 데이터는 분석 및 처리가 가장 큰 이슈이다. 따라서 단순한 물리적인 크기가 아닌 데이터의 어떤 속성에 따라 중요성을 판단하고 그것을 처리하는데 어려움이 있느냐 없느냐인 것이다.

둘째, 데이터의 속도(Velocity): 데이터를 처리하는 속도이다. 정의 부분에서도 설명했듯이 배치 분석만을 의미하는 것이 아니다. 필요에 따라서 수 많은 사용자 요청을 실시간으로 처리한 후 처리 결과를 보내주는 기능도 필요하다.

셋째, 데이터의 형태(Variety): 전통적인 기업의 데이터 분석은 기업 내부에서 발생하는 운영 데이터인 ERP, SCM, MES, CRM 등의 시스템에 저장되어 있는 RDBMS 기반의 정형

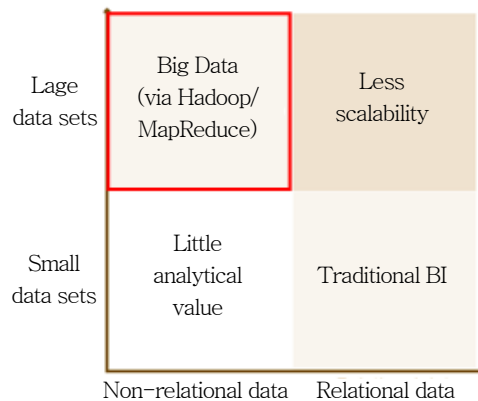
데이터였다. 이러한 정형 데이터는 잘 정제되어 있고 의미도 명확하다. 그리고, 스키마를 포함하는 XML, HTML 등의 반정형 데이터도 있다. 하지만, 최근에는 이런 데이터뿐만 아니라 기업 외부에서 발생하는 SNS, 블로그, 뉴스, 게시판 등의 데이터나 사용자가 업로드한 파일, 콜 센터의 고객 상담 내용 등의 비정형 데이터도 처리해야 한다.

(그림 1)은 2011년 TDWI Research에서 발표한 빅 데이터의 3대 요소를 나타내었다.

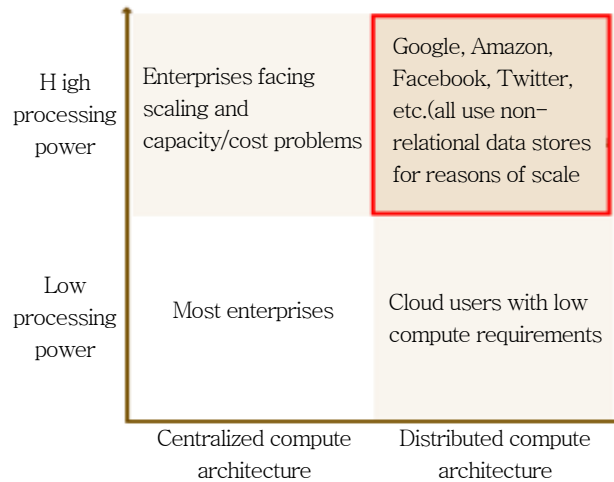


(그림 1) 빅 데이터의 3대 요소

(그림 2)는 2010년 PWC에서 발표한 데이터 특성이 형태와 크기에 따라 구분한 그래프이다. 빅 데이터의 포지션을 잘 나타내주고 있다[4].



(그림 2) 데이터 특성에 따른 빅 데이터의 포지션



(그림 3) 프로세서 파워에 따른 컴퓨터구조의 포지션

(그림 3)은 데이터 아키텍처 배치를 컴퓨팅 아키텍처에 따른 프로세싱 파워에 따라 구분한 그래프이다. 현재 빅 데이터의 위치를 잘 표현한 그래프이다.

3. 빅 데이터의 분석 기술

지금까지의 데이터 분석 기술은 대부분 한대의 컴퓨터상의 메모리, 파일시스템, 데이터베이스에 데이터를 저장하고 이를 기반으로 데이터를 분석하는 알고리즘을 실행하는 구조였다. 대부분의 통계 툴들은 여전히 메모리에 데이터를 로딩해서 통계/분석/마이닝 알고리즘을 실행하는 것이 기본구조이다. 이것이 데이터베이스 시스템이 나오면서 대용량의 데이터를 처리할 수 있어 규모가 커지게 되었다. 하지만 여전히 이러한 분석 시스템의 구조는 싱글머신/싱글코어에 최적화되어 있었으며, 최근에는 싱글머신/멀티코어에서 실행할 수 있는 다양한 알고리즘의 개발과 시스템들이 등장하고 상용화되어 쓰고 있다. 지금까지 빅 데이터라고 하는 것을 처리하기 위해서는 몇백 기가 메모리와 SAN 스토리지로 대용량의 파일시스템을 마운트할 수 있는 고사양 고가의 하이엔드급 서버를 이용해서 DW, DM 을 구축해 왔다. 데이터 증가에 따른 시스템 확장은 더 고사양의 장비로 교체하거나 CPU/ 메모리/ 디스크 증설이라는 방식을 이용해서 하는 scale-up 방식만이 유일했다. 문제는 최근 구글, 아마존, 야후, 페이스북, 트위터와 같은 인터넷 기업들이 고객들의 사용 로그와 트랜잭션 로그를 기반으로 데이터 마이닝과 이를 기반으로 하는 서비스, 광고 플랫폼을 구축하고자

하면서 그 한계에 이르게 되었다. Terabyte 에서 Petabyte 규모의 데이터를 분석하여 검색 엔진, 소셜 서비스, 광고 등을 하기에는 기존의 시스템, 소프트웨어 아키텍처로는 불가능했던 것이다. 뿐만 아니라 이들이 처리해야 하는 데이터들은 데이터베이스에 깔끔히 정리된 정형화된 데이터가 아니라, 웹을 통해서 수집한 다양한 비정형 데이터와 함께 비디오, 사진, 음향 등 다양한 미디어 정보를 수집해서 분석해야 하기 때문에 더욱 힘들어질 수밖에 없게 된 것이다. 구글은 이러한 문제점을 해결하기 위해 MapReduce 라는 프로그래밍 모델과 대용량 분산처리 프레임워크와 대용량 데이터를 효과적으로 저장하고 확장할 수 있는 구글 파일시스템(GFS) 기술을 활용하였다.

구글이 가진 기술을 참고해서 등장한 다양한 MapReduce 프레임워크 중에서 가장 주목을 받고 그 기반으로 커다란 에코시스템을 갖추게 된 것이 바로 자바 기반의 Apache Hadoop 이다. Hadoop 은 다수의 서버를 묶어 분산 처리하는 플랫폼이며, 분산처리하는 ‘Map’ 단계와 결과를 취합하는 ‘Reduce’ 단계로 이루어진 구글의 MapReduce 모델을 본떠 만든 것으로 야후의 더그커킹에 의해서 처음으로 개발되고 배포되었다. Hadoop 은 예전 리눅스의 등장으로 OS 시장이 틀을 크게 바꾸었듯이 빅 데이터 분석 시장에 있어서 커다란 대안으로 등장하고 있다. 야후 내부에서 사용하던 이 기술이 오픈 소스로 발표되어 크게 주목을 받으면서 사실상 현재 페이스북, 트위터, 링크드인, 이베이, 아마존 등 많은 글로벌 인터넷, 커머스 업체들은 빅 데이터 처리를 위해서 Hadoop 의 사용을 당연시 하고 있으며, 이를 기반으로 한 다양한 처리 프레임워크나 기술들을 공개하고 있고 그 저변을 매우 빠르게 넓혀가고 있다.

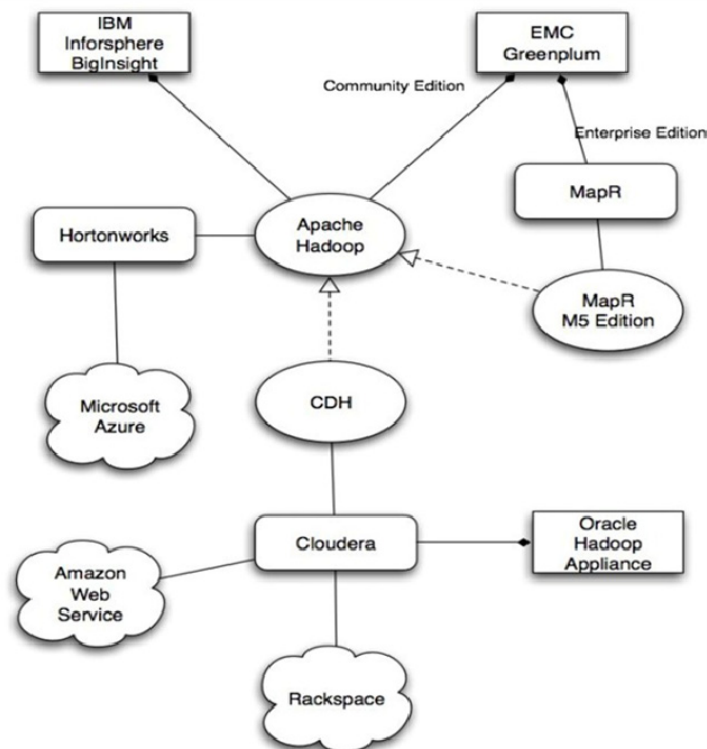
국내의 대표 포털 네이버, 다음 등 인터넷 기업들뿐 아니라 S클라우드를 준비하고 있는 삼성전자와 같은 제조사 역시 스마트폰, 스마트 디바이스를 위한 콘텐츠 서비스와 이를 통해서 발생하는 엄청난 로그 데이터 처리를 위해 Hadoop 을 적극적으로 활용하고 있다. 그리고, 넥스알(NexR)이 Hadoop 및 클라우드 기술을 기반으로 다양한 컨설팅 및 사업을 추진하였고, 2011 년 말 KT 의 자회사로 인수되면 크게 주목을 받았다. 최근에는 KT 이노츠(innotz)와 합병되면서 KT 클라우드웨어라는 회사로 거듭나면서 사업 영역과 규모가 더욱 커졌다. 넥스알은 꾸준히 국내의 Hadoop 오픈소스 커뮤니티의 활동을 적극 지원하고 있고, 최근에는 RHive 라고 하는 R 와 Hive 을 결합한 시스템을 오픈소스로 공개하는 등 국내의 Hadoop 저변 확대에 많은 지원을 하고 있다.

가. 빅 데이터의 솔루션 업체

대표적인 Hadoop 솔루션 업체로 Cloudera 와 Hortonworks 를 들지만, 엔터프라이즈 업체로 MapR 가 있다. 야후에서 분사한 Hortonworks 가 Hadoop 코어의 기술과 아키텍처 개선 등을 담당하고, Cloudera 는 빅 데이터와 클라우드 시장의 기술지원, 교육 및 배포판을 제공하고 있다. 전통적인 솔루션 업체들을 살펴보면, IBM 은 Apache Hadoop 을 기반으로 자신들의 Basic 과 Enterprise 배포판을 가지고 있고, EMC 는 자신의 DW 솔루션인 Greenplum 에 MapR 의 배포판을 통합해서 제공하고 있으며, Apache Hadoop 을 기반으로 Community Edition 을 제공하고 있다. Oracle 은 자신의 하드웨어에 Cloudera 를 결합하여 Hadoop Appliance 를 제공하고 있다.

이밖에 다음과 같이 Apache Hadoop 을 기반으로 빅 데이터 시장에 제품을 소개한 회사가 있다.

- DataStax: Hadoop+ Hive+ Cassandra, <http://www.datastax.com/>



(그림 4) Hadoop 솔루션업체와의 관계도

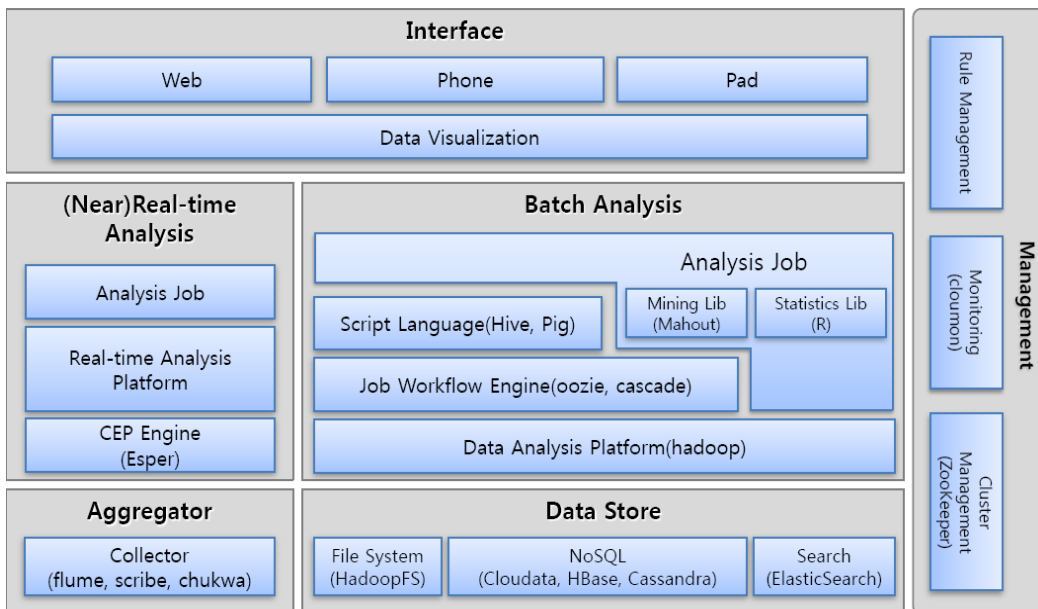
- Datameer: Analytic Solution, <http://datameer.com>
- Hadapt: Analytics Platform, <http://www.hadapt.com/>
- HStreaming: Real-time processing, <http://hstreaming.com/>

(그림 4)에 Hadoop 솔루션 업체와 기존의 전통적인 BI, DW 등의 관계를 간단하게 정리하였다.

나. 빅 데이터 기술

빅 데이터를 다루는 기술들은 어떤 것들이 있을까? 빅 데이터라는 용어를 이끌어 낸 것도 Hadoop 과 NoSQL 의 성공에 있다고 볼 수 있기 때문에 가장 중요한 기술은 Hadoop 이라 할 수 있다. Hadoop 자체는 파일 시스템과 분산 처리 플랫폼이지만 Hadoop 을 중심으로 다양한 에코 시스템이 구축되면서 이제 Hadoop 은 빅 데이터에 있어 산업계 표준이라고 할 수 있다.

Hadoop 이 빅 데이터를 분석 처리하는데 주로 사용될 것으로 생각하지만, 비정형 데이터뿐만 아니라 정형 데이터도 처리할 수 있어야 하기 때문에 다양한 기술들이 필요하다. 이러한 다양한 요소 기술들의 집합체를 Hadoop ECOsystem 이라고 한다. (그림 5)는 Hadoop



(그림 5) Hadoop ECO system 의 software stack

ECOsysteem의 software stack 이다[5]. Data Store 부터 인터페이스까지 6 개 파트로 구분된다.

이런 기술들이 필요에 따라 적절하게 도입되어야 빅 데이터를 처리할 수 있는 시스템을 구축할 수 있다. 언급한 기술 하나하나 쉽지 않은 기술이며 아직 성숙되지 않은 기술도 많다. 다행인 것은 대부분 오픈 소스로 코드와 기술이 많이 공개되어 있다는 것이다.

여러 기술이 있지만 이 중 가장 중요한 기술은 어떤 데이터를 분석할 것인가를 정의하고, 데이터 간의 관계를 찾아서 의미 없는 데이터로부터 의미를 찾아내는 기술이 가장 중요한 기술이다. <표 1>에 Hadoop ECOsysteem의 요소기술들을 정리하였다[6].

최근 들어 MySQL, Oracle, DB2, and SQL Server 등의 상용 관계형 DB 보다 성능이 우수한 오픈소스 및 인덱스 기반의 데이터 저장 구조를 가진 이른바 NoSQL(Not only SQL)이라는 새로운 형태의 데이터 스토리지 솔루션이 개발되었다.

재정이 빈약한 Web 2.0 기업들이 구글이나 아마존, MS 등의 대기업군을 따라 잡으려

<표 1> Hadoop ECOsysteem의 요소기술

Scope	Project	Status
Structured data	HBase, Cassandra, MongoDB, CouchBase, Cloudata, Riak	Hadoop subproject
Storage, Parallel processing	Hadoop	Top level project, healthy community
Parallel machine learning	Mahout, Radoop, Rapid Miner	Incubating
Distributed Queue	Kafka, Bookeper	stable
Enterprise search	Solr, Katta, Elastic search	stable
Metadata extraction	Tika	Incubating
Free text indexing and search engine	Lucene	Stable, de-facto standard solution
Crawler and search engine	Nutch	Stable, slow process
Parallel scripting language	Hive, Pig, Avro	Incubating, contributed by Yahoo
Distributed locking naming service(coordinator)	ZooKeeper, Chukwa	Hosted on SourceForge, developed by Yahoo
Unstructured information	UIMA	Incubating, contributed by IBM
MapReduce extension	TupleFlow	non Apache, BSD license
MapReduce extension	cascading	non Apache, GPLv3.0 license
Statistics, Matrix	R, RHIPE, CUDA, Segue	stable
Streaming Analysis	S4, Esper, Akka, Storm	stable
Distributed Cache	Redis, Membase	stable
Graph Analysis	Hama, GoldenORB, Pregel	stable

면 오픈 소스 기반의 NoSQL 밖에는 대안이 없다. 한 예로, 페이스북은 기존 데이터베이스인 MySQL이 아니라 카산드라(Cassandra) 데이터 스토어를 개발하였다.

그래서 Rick Cattell의 High Performance Scalable Data Stores라는 Paper 내 언급된 NoSQL이라는 데이터 저장 시스템의 공통된 특징들을 아래와 같이 정리하였다[7].

- Key & Value 로 저장
- 분산 환경 지원
- Call level interface 지원(DBMS에 접근하는 표준)
- 막대한 양의 데이터를 처리할 수 있는 대용량 데이터의 빠른 인덱싱
- 클러스터나 그리드에서의 구동을 위해 다양한 테이블로 데이터베이스를 나눠야 하는 복잡한 작업과 ‘샤딩(sharding)’ 없이 손쉽게 저렴하게 여러 서버들의 수평적 확장(horizontal scaling)됨
- 데이터의 스키마와 속성들을 동적으로 정의

4. 빅 데이터의 처리 기술

빅 데이터 처리 기술은 다음 3가지의 기술로 요약되며, 그 활용 분야를 예로 들었다.

가. 분석 기술

데이터를 분석하는 기술과 방법을 의미하며 통계, 데이터마이닝, 기계학습, 자연어 처리, 패턴인식 등이 이에 해당된다. 가장 기본적인 오픈소스 기반 빅 데이터 처리기술이 Hadoop이다. 분산된 파일을 처리하는데 최적화된 Hadoop은 저장장치를 병렬로 연결시켜 데이터를 처리하는 방식으로, 별도의 스토리지가 필요 없고, 대규모 데이터를 빠르게, 비용 효율적으로 처리하는데 적합하다. 비구조화된 데이터를 구조화하는 기술이 MapReduce이고, SQL을 사용하지 않고, 대량의 데이터를 빠르게 처리하는 기술이 NoSQL이다.

이러한 기술들은 데이터의 안정성이 정합성보다 속도, 비용에 초점을 맞췄기 때문에 일부 데이터의 유·손실이 발생해도 처리 결과가 영향을 받지 않는 업무에 적합하다. 로그 데이터 처리 혹은 분석하는 업무를 대표적인 예로 들 수 있다.

나. 표현 기술

일반적으로 데이터의 시각화로 알려져 있으며, 분석된 데이터의 특징이나 의미를 쉽게

알 수 있도록 잘 표현해주는 기술이다. 대표적인 시각화 표현 기술은 R로 통계 계산 및 시각화를 위한 언어 및 개발환경을 제공하며, 기본적인 통계 기법으로부터 모델링, 최신 데이터 마이닝, 시뮬레이션, 수치해석 기법까지 구현 가능하다. 구현 결과는 그래프 등으로 시각화 할 수 있으며, 다른 프로그래밍 언어와 연결도 용이하다. 계몽, 신약연구와 금융 예측 분석에 활용되고 있다.

다. 분석 인프라

분석과 표현을 수행할 수 있도록 해주는 기반기술과 플랫폼들이라고 할 수 있으며, 이러한 분석 인프라는 다시 대규모 데이터를 안정적으로 수집해서 저장하는 기술, 저장된 것을 효과적이면서도 빠르게 처리할 수 있는 기술, 저장된 데이터를 다양한 방식과 용도로 사용할 수 있도록 가공하고 관리해주는 기술 등으로 나누어 진다. 예를 들면 BI(Business Intelligence), DW(Data Warehousing), Cloudcomputing, 분산 데이터베이스(NoSQL), 분산 병렬처리(Hadoop MapReduce), 분산 파일시스템 등이 분산 인프라에 속하는 기술들이다.

5. 빅 데이터 미래 기술

빅 데이터의 미래 기술을 다음 두 가지로 정리해 보았다.

가. 실시간 빅 데이터 분석 기술

여기서 말하는 실시간의 의미는 디바이스에서 말하는 하드웨어 레벨의 실시간 데이터 프로세싱이 아니라, 비즈니스 레벨 또는 서비스 레벨에서의 실시간 데이터 분석기술이다. 예를 들어서 새로운 광고를 웹사이트에 노출시켰을 때 방문자들의 클릭 스트림을 얼마나 빨리 처리해서 고객들의 반응을 분석하고 리포팅하는 것, 그리고 엄청나게 폭주하는 주식 거래의 실시간 트랜잭션을 분석하는 것도 포함된다. 이러한 실시간 데이터 분석을 위해서 주목 받는 기술 중 하나가 Complex Event Processing(CEP)라고 하는 기술이다. 다시 말해 실시간으로 발생하는 복수의 이벤트로부터 특정 패턴을 찾아내서 원하는 데이터 처리나 알림 서비스를 하는 기술이다.

나. 실시간 빅 데이터 형상화(Visualization) 기술

소셜 네트워크 및 서버 등에서 발생하는 다양한 데이터 로그를 한데 모아 분석하고, 이

를 형상화하여 IT 자원의 효율성을 높이는 기술이다. 분석된 결과를 사용자가 이해하기 쉬운 형태로 효과적으로 데이터를 보여줄 수 있는 표현방식과 이를 프로세싱하기 위한 알고리즘 등의 기술이다.

6. 결론

본 고에서는 빅 데이터 시대에 필요한 다양한 기술들에 대해서 살펴 보았다. 빅 데이터 분석을 통해 기업은 사회와 인류의 유용한 정보를 효과적으로 분석해서 전략 마케팅을 펼치고, 정부는 사회구성원들이 쏟아내는 막대한 정보를 분석을 통해 교통, 세금, 범죄, 재난 대처 등 사회 각 영역에서 과거와는 차원이 다른 정확한 정보를 토대로 예산 집행의 효율성과 공공 서비스의 질을 효과적으로 높여 빅 데이터 시대의 도래에 따른 국가 경쟁력 향상을 위한 준비가 필요하다.

마지막으로, 정부/기업 양측의 빅 데이터 활용의 중요성을 공감하고, 빅 데이터 활용 방안 및 전문 인력 양성을 통해 미래사회 효과적 대비를 위한 역할 분배 논의가 요구된다.

- PB(Peta byte): 디지털 신호의 처리 속도 또는 용량을 표시하는 단위로 10¹⁵ 이다. 1PB 는 1,024TB 이다.
- ERP(enterprise resources planning): 전사적 자원관리, 기업활동을 위해 사용되는 기업 내의 모든 인적, 물적 자원을 효율적으로 관리하여 궁극적으로 기업의 경쟁력을 강화시켜 주는 역할을 하는 통합정보시스템이다.
- SCM(Supply Chain Management)은 기업에서 생산·유통 등 모든 공급망 단계를 최적화해 수요자가 원하는 제품을 원하는 시간과 장소에 제공하는 ‘공급망 관리’를 뜻한다.
- CRM(customer relationship management) 고객 관계 관리는 기업이 고객 관계를 관리해 나가기 위해 필요한 방법론이나 소프트웨어 등을 가리키는 용어이다. 기업들이 고객들의 성향과 욕구를 미리 파악해 이를 충족시켜 주고 기업들이 목표로 하는 수익이나 광고효과 등 원하는 바를 얻어내는 기법을 말한다.
- BI(Business Intelligence)는 데이터를 정보화시키고 이를 이용하여 회사 경영에 도움을 주는 정보를 추출 기법을 말한다.
- DW(Data Warehouse)는 기존 시스템의 데이터베이스에 축적된 데이터를 추출하여

공통 형식으로 변환, 일원화시켜 새롭게 생성된 데이터베이스로 다양한 형태의 데이터에 대해 사용자 분석이 용이하도록 주제지향적, 비휘발적으로 구성된 통합자료 저장소이다.

- DM(data mining) 많은 데이터 가운데 숨겨져 있는 유용한 상관관계를 발견하여 미래에 실행 가능한 정보를 추출해 내고 의사 결정에 이용하는 과정을 말한다.
- NoSQL(Not Only SQL)은 오픈 소스 기반의 고성능의 분산 데이터 저장 시스템으로 관계형 데이터베이스의 한계를 극복하기 위한 데이터 저장소의 새로운 형태의 수평적 확장을 특징으로 한다. 관계형이 아니기 때문에 join 이 없고 고정된 스키마를 갖지 않는다.
- DW 는 운영데이터의 추출, 변환, 적재된 데이터들의 모임으로 큰 DB 라고 생각하면 편하겠다.
- 샤딩(sharding)은 여러 노드에서 테이블 구조를 복제하지만 이러한 노드 사이에서 데이터를 논리적으로 나누는 파티셔닝의 한 양식이다.
- DB2(Database 2) 1983 년에 발표된 미국 IBM 의 관계형 데이터베이스 관리 시스템이다. 여러 사용자들이 여러 개의 데이터 베이스에 동시에 접근할 수 있는 대형 데이터베이스를 위한 시스템이다.

<참 고 문 헌>

- [1] “Big Data: The next frontier for innovation, competition, and productivity”, McKinsey. 2011. 5.
- [2] IDC IVIEW Extracting Value from Chaos June 2011 By John Gantz and David Reinsel.
- [3] TDWI Research 2011 Big Data Analytic Report
- [4] Technology forecast Making sense of Big Data A quarterly journal 2010, Issue 3
- [5] 김형준, Cloud Computing 기술을 활용한 BigData 를 위한 아키텍처 및 기술 2011. 12.
- [6] <http://www.lifeyun.com/hadoop-mapreduce.html>
- [7] High Performance Scalable Data Stores, Rick Cattell. February 22, 2010.