# 빅 데이터 (Big Data)

글 _ 문두환 _ 경북대학교 정밀기계공학과 _ dhmun@knu.ac.kr, 도남철 _ 경상대학교 산업시스템공학부 _ dnc@gnu.ac.kr

## 1. 빅 데이터의 개념

IT 기술에서 빅 데이터(Big Data)는 구조가 복잡하고 데이터 양이 많아서 기존의 데이터베이스 관리 도구로 처리하기 어려운 데이터 집합을 말하는 용어이다[3]. 예를 들어 웹 로그, RFID(Radio Frequency Identification), 센서 네트워크, 소셜 네트워크, 인터넷 검색 인덱싱, 천문학/대기과학 등의 학문적 연구 기록, 전자상거래 이력 등이 바로 그것이다[1].

IT 기술의 발전에 따라 데이터 처리 기술은 급격한 발전을 이루었다. 예를 들어, 10년 전에는 슈퍼 컴퓨터를 이용해도 많은 시간이 걸렸던 작업들을 일반 컴퓨터를 이용하여 짧은 시간 안에 저비용으로 할 수 있게 됐다. 그러나 데이터의 증가 속도는 데이터 저장 및 분석 기술보다 더 급격하게 증가하고 있다. 특히 다양한 센서가 내장된 같은 임베디드 시스템 및 인터넷 환경에서 파일, 이메일, 동영상 등과 같이 구조화되지 않은 비정형 정보들이 폭발적으로 증가할 것이다.

빅 데이터는 단순한 데이터 양의 증가가 아니라 데이터의 형식, 입출력 속도 등을 함께 아우르는 의미다. 빅 데이터를 규정하는 세 축은 다양성 (Variety), 규모 (Volume), 속도 (Velocity)이다[2]. IBM은 세가지 요소 가운데 두 가지를 충족할 수 있으면 빅 데이터 기술이라고 정의한다.

## 2. 빅 데이터의 구성 기술

빅 데이터는 정형화 정도에 따라서 고정형 필드에 저장된 정형 데이터, 고정된 필드는 아니지만 스키마를 포함하는 반 정형 데이터, 고정된 필드에 저장되어 있지 않은 비정형 데이터로 구분된다.

전통적인 OLTP (Online Transaction Processing)나 OLAP (Online Analytical Processing) 기술은 구조화되고 정형화된 데이터를 처리하는데 집중해 왔다. 그러나 현대는 실시간으로 데이터를 수집하고 분석하는 것과 함께 비정형 데이터에서 가치 있는 정보 패턴을 찾는 것이 중요하여 새로운 데이터 관리, 처리, 그리고 분석 기술이 요구된다[1].

빅 데이터를 위한 주요 기술 구성은 인프라 기술, 분석 기법, 표현 기술 등 다음과 같이 세가지로 분류된다[1].

- 인프라 기술: BI (Business Intelligence), 클라우드 컴퓨팅, 분산 데이터베이스, 분산 병렬처리, 분산 파일 시스템
- 분석 기술: 데이터를 분석하는 기술과 방법론 (통계, 데이터 마이닝, 기계 학습, 자연어 처리,

패턴 인식, 예측 모델링)

- 표현 기술: 일반적으로 데이터 시각화로 알려져 있으며 분석된 결과를 보여 주는 기술

빅 데이터와 관련된 최근의 주요 기술적 이슈는 빅 데이터 관련 서비스 (데이터 분석 서비스, 데이터 가시화 서비스), Crowdsourcing model, 빅 데이터를 위한 데이터베이스, 데이터 분석 라이브러리의 구축이다[6].

## 3. 빅 데이터 지원 도구

빅 데이터 인프라로 대표적인 기술이 분산 데이터 처리 플랫폼[4]이며, Yahoo에 적용된 Hadoop, Facebook에 적용된 Cassandra, Zynga에 적용된 Membase [6] 등이 있다. 본 기사에서는 세가지 플랫폼 중에서 Hadoop을 설명한다.

Hadoop은 분산 파일 시스템인 HDFS (Hadoop Distributed File System)와 데이터베이스 시스템인 Hbase, 분산 프로그래밍 프레임워크인 MapReduce를 기반으로 데이터 수집 시스템인 Chukwa나 Flume, 또는 대용량 데이터 패턴을 분석하는 기계 학습 프레임워크인 Mahout 등으로 구성되어 있다.

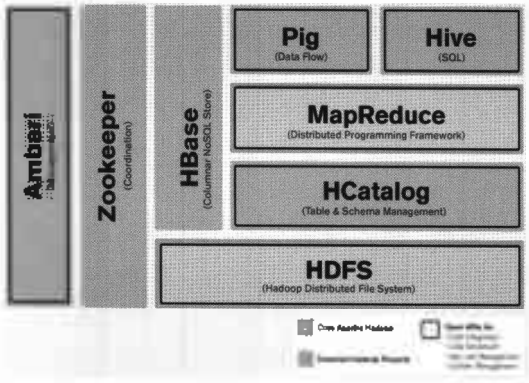


그림 1. Hadoop의 구성 요소 [1]

Hadoop 플랫폼을 활용한 빅 데이터 분석 과정의 예가 다음 그림이다. 그림에서 엔지니어링팀과 세일즈팀의 반구조화 데이터와 소셜 미디어 등의 비정형 데이터는 Hadoop에서 통합이 된 후에 BI 도구를 활용하여 연관관계를 분석을 하여 의미 있는 정보를 만들게 된다. Hadoop과 기존 RDBMS의 차이점은 분산 파일 시스템의 사용에 있으며, MapReduce를 활용하여 분산되어 있는 파일을 병렬 처리함으로써 대용량 데이터의 처리가 가능하다.

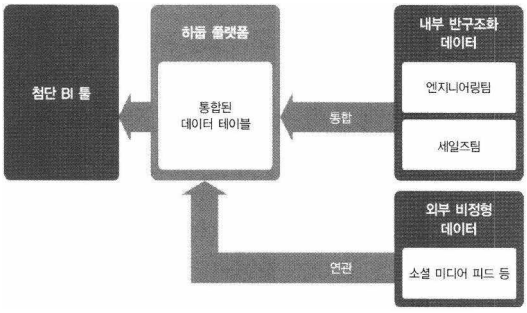


그림 2. 빅 데이터의 분석 과정 [2]

분산 데이터 처리 플랫폼 이외의 빅 데이터 인프라 기술로 NoSQL 데이터베이스가 있다. NoSQL 데이터베이스는 전통적인 관계형 DBMS(Database Management System)와는 달리 수평적 데이터 확장에 강점을 있어 대용량 데이터를 유연하게 처리할 수 있다[5]. 위에서 설명한 Hbase가 NoSQL 데이터베이스 중 하나이다. Hbase는 구조화된 대용량 데이터의 관리를 위해 Google사가 개발한 분산 데이터베이스로 BigTable의 공개형 버전이다.

빅 데이터는 기존 데이터와 달리 처리의 복잡도가 높고 처리할 데이터의 양이 방대하며 비정형 데이터의 비중이 높다는 점이 특징이다. 빅 데이터의 분석 기법으로는 다음이 있다[1].

- 텍스트 마이닝: 자연어 처리 기술을 기반으로 비정형 텍스트 데이터에서 정보를 찾아내는 기술
- 평판 분석 (Opinion Mining): 웹 상의 데이터로부터 소비자의 의견을 수집하여 분석하는 기술

- 소셜 네트워크 분석: 개인 또는 그룹의 소셜 네트워크 내 영향력/관심사/성향 및 패턴을 분석하는 기술
- 클러스터 분석: 다양한 데이터 간의 유사도를 정의하고 각 데이터 간의 거리를 구하여 가까운 거리에 있는 데이터들을 병합하는 기술
- R: 통계 계산 및 시각화를 위한 분석 엔진

## 4. 빅 데이터의 PLM 시스템 응용

PLM (Product Lifecycle Management) 시스템 관점에서 빅 데이터의 응용 분야로는 다음이 있다.

- 고객 요구 분석: 빅 데이터 기술을 통하여 다양한 고객 요구 정보를 확보, 분석하여 제품 사양을 도출할 수 있는 시스템을 개발
- 설계, 생산, 고객 지원 자료의 통합: 설계, 생산, 그리고 고객 지원 업무는 각각 PLM, ERP(Enterprise Resource Planning), SCM(Supply Chain Management), 그리고 CRM(Customer Relationship Management) 등의 특화된 시스템 상에서 이뤄지고 있으나 빅 데이터 기술을 기업 차원의 광범위한 제품 정보를 통합하는데 사용
- 제품 개발 지식 관리: PLM 시스템에는 방대한 제품 개발 자료가 축적되어 있어 빅 데이터 기술을 적용하여 제품 개발에 의미 있는 다양한 지식들을 발견하여 활용
- 제품 개발 과정 관리: 대표적인 비정형적 업무 중의 하나로 알려져 있는 제품 개발 과정 중에 발생하는 많은 양의 정보를 수집 분석하여 제품 개발 과정의 평가와 분석을 수행

현재 상용 PLM 시스템에 적용된 빅 데이터 관련 도구로는 Siemens PLM Software의 Active Workspace[7]와 Dassault Systemes의 Netvibes[8] 등이 있다. Active Workspace는 전체 PLM 시스템에 접근하기 위한 개인화된 환경으로 제품 데이터 시각화 및 탐색, 제품 정보 비교 및 보고, 컨텍스트 구성 및 공유 기능 등을 제공한다. Netvibes는 실시간 모니터링, 소셜 분석, 지식 공유, 그리고 의사결정을 위한 개인화된 환경(dashboard)를 쉽고 바르게 생성할 수 있는 기능을 제공하며 기업 정보 시스템 및 여러 디바이스와의 연동이 가능하다.

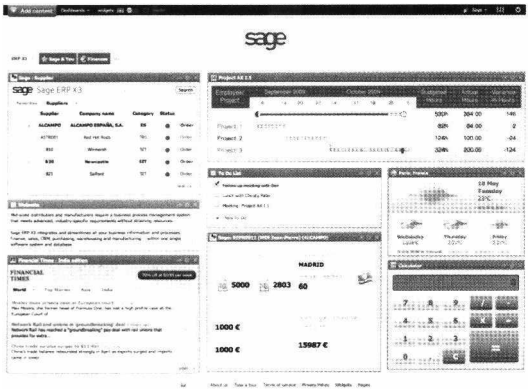


그림 3. 기업 차원의 Netvibes의 적용 사례 [9]

## 5. 결론

이 기사에서는 참고문헌을 바탕으로 빅 데이터 개념, 구성 기술, 지원 도구에 대해서 살펴보았다. 지금까지는 빅 데이터 기술이 주로 비즈니스 영역에 집중되어 연구되었지만 앞으로 엔지니어링 정보 시스템에서 빅 데이터 기술을 어떻게 활용할지 고민해야 할 시점이다. 특히 PLM 시스템 관점에서 보면, PLM 데이터베이스가 정형화된 데이터 저장소 및 기본적인 데이터 처리 기능을 제공하기 때문에 엔지니어링 데이터를 분석하여 유의미한 정보를 찾아서 활용하는 PLM 기반 데이터 분석 기술의 연구가 중요할 것으로 판단된다.

참고문헌

1. IDG Korea, "빅 데이터의 이해", IDG Tech Report, http://www.itworld.co.kr/techlibrary, 2012.06.28.
2. 하정욱, "클라우드와 빅 데이터", IDG Summary, http://www.itworld.co.kr/techlibrary, 2012.06.28.
3. Wikipedia, http://en.wikipedia.org/wiki/Big_data, 2012.06.28.
4. Tom White, "Hadoop: The Definitive Guide", 1st Edition, O'Reilly Media, 2009.
5. Dorin Carstoiu, Elena Lepadatu, Mihai Gaspar, "Hbase - non SQL Database, Performances Evaluation", International Journal of Advancements in Computing Technology, Vol. 2, No. 5, pp. 42-52, 2010.
6. Munish K Gupta, "Bigdata Trends in 2012", http://www.techspot.co.in/2012/01/bigdata-trends-in-2012.html, 2012.06.28.
7. Siemens PLM Software Active Workspace, http://www.plm.automation.siemens.com/ko_kr/products/teamcenter/active-workspace/index.shtml?stc=kriia420175, 2012.07.02.
8. Netvibes, http://www.netvibes.com/ko-kr, 2012.07.02.
9. Sage releases Sage Enterprise WebTop, powered by Netvibes, http://kingkool68.com/items/view/26538/sage-releases-sage-enterprise-webtop-powered-by-netvibes, 2012.07.02.

※ [작성자 주] 이 기사는 상기의 참고문헌을 바탕으로 작성된 글임을 밝힙니다.