빅데이터 처리 프로세스 및 활용

이성훈*, 이동우**

Big Data Processing and Utilization

Seong-Hoon Lee*, Dong-Woo Lee**

요 약 우리사회는 점점 더 융/복합 현상이 가속화되고, 광범위한 영역으로 확대되고 있다. 이러한 중심축에는 정보통신 기술이 자리잡고 있음은 당연한 일이다. 일례로 정보통신기술과 의료산업의 융합의 결과로 스마트 헬스케어 산업이 등장하였으며, 모든 분야에 정보통신 기술을 접목하고자 하는 노력들이 계속되고 있다. 이로 인해 우리주변에는 수많은 디지털 데이터들이 만들어지고 있다. 또 다른 한편으로는 대중화 되고 있는 스마트폰, 태블릿PC와 카메라, 게임기기등을 통하여 다양한 데이터들이 생성되고 있다. 본 연구에서는 광범위하게 발생하고 있는 빅데이터에 대한 활용 상태를 알아보고 빅데이터 플랫폼의 한 축인 처리 프로세스들에 대해 비교, 분석하였다.

주제어 : 빅데이터, 처리 프로세스, 융/복합. 빅데이터 활용

Abstract Our society has two prospective properties because of IT technology. Firstly, it is accelerated a degree of convergence. And convergence regions are expanded. For example, smart healthcare region was created by IT technology and medical industry. The efforts to convergence will be continued. Because of these properties, A number of data are made in our life. Through many devices such as smart phone, camera, game machine, tablet pc, various data types are produced. In this paper, we described utilization of Big Data. And we analysed Big Data processing process.

Key Words : Big Data, Processing, Convergence, Big Data Utilization

1. 서론

“오늘 태어난 아이가 평생 동안 살아가면서 만들어 낼 데이터의 양은 미국 의회 도서관에 보관된 데이터의 약 70배에 이를 것이다.” “하나의 정보가 저장될 때 저장되지 못하고 있는 정보는 100개에 이른다.” “60초당 하나씩 유튜브의 동영상이 업로드되고 있다.”

위에서 기술된 내용들은 빅데이터의 등장을 표현하기 위해 등장한 수식어들이다.

오늘날 정보통신 분야에서의 화두는 단연 빅데이터 및 클라우드 컴퓨팅, 융복합을 이야기할 수 있을 것이다. 리서치 자문기업인 가트너 역시 최근에 모바일 기기 전쟁, 전략적 빅 데이터 등 2013년 기업들이 전략적으로 대응해야 하는 10대 기술 및 트렌드를 발표하였다. 가트너는 2013년에는 모바일폰이 전세계에서 가장 널리 사용되는 웹 액세스 기기로서 PC를 추월하게 될 것이며, 2015년에 이르면 선진국 시장에서 판매된 휴대폰의 80% 이상을 스마트폰이 차지하게 될 것으로 예측하고 있다[3].

또 개인이 자신의 개인적인 콘텐츠를 보관하고, 자신의 서비스와 선호하는 대상에 접근하며 자신의 디지털 생활을 집중시키는 장소는 PC에서 퍼스널 클라우드로 점차 대체될 것으로 예측했다.

빅데이터라는 용어가 처음 등장했을 때에는 그 의미를 각기 다르게 해석하였다. 어떤 그룹에서는 “테라바이트 이상의 데이터”라고 정의하였고 또 다르게는 “대용량 데이터를 처리하는 아키텍처”라고 정의하기도 하였다. 하지만 “빅”이라는 단어 자체가 상대적 의미를 지니기 때문에 기준이 되는 데이터 용량을 정의하기엔 적절하지

*백석대학교 정보통신학부 교수
**우송대학교 컴퓨터정보학과 교수, 교신저자.
논문접수: 2013년 3월 14일, 1차 수정을 거쳐, 심사완료: 2013년 4월 15일, 확정일: 2013년 4월 20일

못한 것 같다.

빅데이터는 기존 데이터에 비해 너무 방대하여 기존의 방법이나 도구로 수집, 저장, 분석등이 어려운 정형 및 비정형 데이터들을 의미한다. 세계적인 컨설팅 기관인 Mckinsey지는 2011년 한 보고서에서[5] 빅데이터의 정의는 "기존 데이터베이스 관리도구의 데이터 수집, 저장, 관리, 분석하는 역량을 넘어서는 데이터 셋 규모로서 그 정의는 주관적이며 앞으로도 계속 변화될 것이다"라고 언급하고 있다.

기존의 전통적인 데이터 개념과 현재 화두가 되고 있는 빅데이터의 특성을 비교하면 다음 <표 1>과 같다.

〈표 1〉 기존데이터/빅데이터 비교

기존데이터	빅데이터
Gigabytes to Terabytes	Petabytes to Exabytes
Centralized	Distributed
Structured	Semi-structured and Unstructured
Stable data model	Flat schemes
Known complex interrelationships	Few complex interrelationships

빅데이터가 갖추어야 하는 요소기술로서 미디어관련 데이터 크기(Volume), 데이터 입/출력 속도(Velocity), 데이터 형태(Variety)가 있다. 크기는 일반적으로 수십 테라 혹은 수십 페타 바이트 이상 규모의 데이터 속성을 의미한다. 속도는 대용량의 데이터를 빠르게 처리하고 분석할 수 있는 속성이다. 융복합 환경에서 디지털 데이터는 매우 빠른 속도로 생산되므로 이를 실시간으로 저장, 유통, 수집, 분석처리가 가능한 성능을 의미한다. 형태는 다양한 종류의 데이터를 의미하며, 정형화의 종류에 따라 정형(Structured), 반정형(Semi-Structured), 비정형(Unstructured)으로 분류될 수 있다[2].

오늘날 전 세계에서 다루어지는 디지털 정보량은 2년에 2배씩 증가하고 있다고 한다[4][6][7]. 정보통신 기술이 다른 산업들과 융복합되면서 방대한 량의 데이터들이 생성되고 있는 가운데 사회변화에 따른 삶의 질에 대한 욕구 및 현안 해결에 빅데이터들의 활용이 매우 중요한 과제로 떠오르고 있는 것이다.

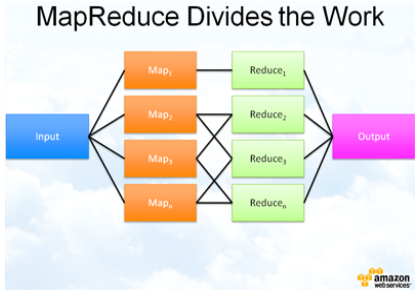
2. 빅데이터 플랫폼

빅데이터를 다루는 여러 플랫폼을 저장 시스템, 처리 프로세스(방식), 분석 메커니즘이라는 세 가지 측면으로 나누어 볼 수 있다. 본고에서는 위에서 기술된 3대 요소 중 처리 프로세스와 관련된 플랫폼 기술에 대해 기술하였다. 저장 시스템으로 병렬 DBMS와 NoSQL은 모두 대량의 데이터를 저장하기 위해 수평 확장 접근 방식을 취하고 있다는 점에서는 동일하다. 이 이외에도 SAN (Storage Area Network), NAS(Network Attached Storage)와 같이 기존 저장 장치 기술도 있고, Amazon S3나 OpenStack Swift와 같은 클라우드 파일 저장 시스템, GFS(Google File System), HDFS(Hadoop Distributed File System)와 같은 분산 파일 시스템 등이 모두 대량의 데이터를 저장하기 위한 기술이다.

빅데이터를 다루는 처리 프로세스로서 병렬 처리의 핵심은 분할 점령(Divide and Conquer)이다. 즉 데이터를 독립된 형태로 분할하고 이를 병렬적으로 처리하는 것이다. 빅데이터의 데이터 처리란 이렇게 문제를 여러 개의 작은 연산으로 나누고 이를 취합하여 하나의 결과로 만드는 것을 말한다. 물론 연산 의존성이 있는 경우에는 병렬 연산의 이점을 살릴 수 없다는 문제점이 존재한다. 이러한 점을 고려한 데이터 저장과 데이터 처리가 필요하다.

2.1 Map_Reduce 프로세스

대용량의 데이터를 처리하는 기술 중 가장 널리 알려져 있는 것은 Apache Hadoop과 같은 Map-Reduce 방식의 분산 데이터 처리 프레임워크일 것이다[1][8][9]. Map-Reduce 방식의 데이터 처리 의미는 아래 그림 1과 같으며, 다음과 같은 특징이 있다.



[그림 1] Map-Reduce 원리(출처: 아마존웹서비스)

① 특별한 저장소가 아닌 일반적인 내장 하드 디스크 드라이브를 사용하는 일반 컴퓨터로 연산을 수행한다. 각 컴퓨터는 서로 매우 약한 상관 관계를 가

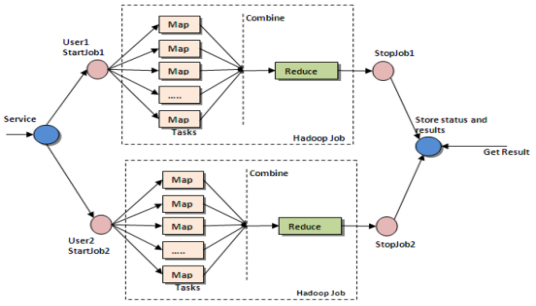
지고 있기 때문에, 수백~수천 대까지 확장 가능.

② 많은 수의 컴퓨터가 처리에 참가하므로, 하드웨어 장애 등의 시스템 장애가 예외적인 상황이라기 보다는 일반적인 상황이라 가정.

③ Map과 Reduce라는 간단하고 추상화된 기본 연산으로 복잡한 여러 문제를 해결할 수 있도록 한다. 병렬 프로그램에 익숙하지 않은 프로그래머라도 쉽게 데이터에 대한 병렬 처리 가능.

④ 많은 수의 프로세서에 의한 높은 처리율(throughput)을 지원.

다음 [그림 2]는 Map-Reduce 방식의 프로그래밍 개념을 나타낸 것이다.

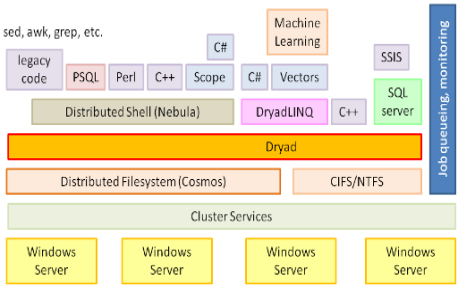


[그림 2] Map-Reduce 프로그래밍 개념

2.2 Dryad 프로세스

Dryad는 프로그램과 프로그램 사이의 데이터 채널을 그래프 형태로 구성해서 병렬 데이터 처리를 할 수 있도록 하는 프레임워크이다.

Map-Reduce 프레임워크를 사용하는 개발자가 할 일은 Map 기능과 Reduce 기능을 작성하는 것이었는데, Dryad를 사용할 경우에 개발자가 할 일은 해당 데이터를 처리하는 그래프를 만들어 내는 것이 된다.



[그림 3] Dryad 소프트웨어 스택

[그림 3]은 Dryad의 소프트웨어 구조를 보이고 있다. Dyrad에서는 DAG(Direct Acyclic Graph) 형태의 데이터 흐름을 처리해줄 수 있게 해주고 있다.

2.3 Apache Pig 프로세스

Map-Reduce나 Dryad와 같은 병렬 데이터 연산 프레임워크는 빅데이터를 처리하기 위한 기능을 충분히 제공하지만, 경험이 많지 않은 개발자나 데이터 분석가, 데이터 마이너가 무리 없이 사용하기에는 어느 정도 진입 장벽이 있는 것이 사실이다. 그렇기 때문에 좀 더 높은 수준의 추상화로 데이터를 더 쉽게 처리할 수 있는 방법이 필요해지게 되었다. 다음에 설명할 Apache Pig와 Apache Hive가 이런 필요에 따라 나온 프레임워크이다.

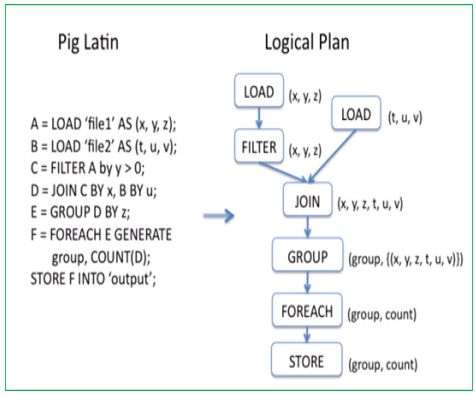
Apache Pig는 고수준의 데이터 처리 구조를 제공하며, 이를 조합해서 대량의 데이터 처리를 가능하게 한다. Apache Pig는 Pig Latin이라는 언어를 지원하는데, Pig Latin은 다음과 같은 특징이 있다.

① int, long, double과 같은 기본 타입 외에 relation, bag, tuple 과 같은 고수준의 구조를 제공.

② FILTER, FOREACH, GROUP, JOIN, LOAD, STORE 등의 관계(relation, table) 연산을 지원.

③ 사용자 지정 함수를 정의할 수 있음.

Pig Latin으로 명세한 데이터 처리 프로그램은 논리적인 실행 계획으로 변환되고, 이것이 다시 Map-Reduce 실행 계획으로 변환되어서 실행된다. 다음 [그림 4]는 Apache Pig의 동작 과정을 나타내는 그림이다.

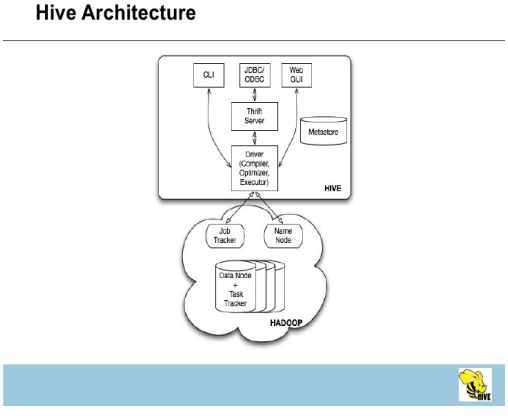


[그림 4] Pig Latin의 Map-Reduce 변환과정

2.4 Apache Hive 프로세스

Apache Pig의 경우는 C나 Java와 같이 절차적 프로그램 언어 형태로 대량의 데이터 처리 프로그램을 작성할 수 있도록 하는 접근 방식이다. 이와 비슷한 접근 방식으로 Google의 Sawzall이 있다. 프로그래밍 언어와 같이 절차적으로 데이터 처리를 명세하지 않고, SQL처럼 선언적 데이터 처리를 할 수 있도록 하는 기술도 있다. 대표적인 것으로 Apache Hive, Google Tenzing, Microsoft의 SCOPE 등이 있다.

Apache Hive는 HDFS나 HBase와 같은 대량의 데이터 원본을 HiveQL이라고 부르는 쿼리 언어로 분석할 수 있는 기술이다. 아키텍처상으로는 Map-Reduce 기반의 실행 부분과 데이터 저장소에 대한 메타데이터 정보와 사용자나 응용프로그램으로부터 쿼리를 받아 실행하는 실행 부분으로 나뉘어져 있다. 사용자에 의한 확장을 지원하기 위해서 scalar 값, aggregation, table 수준에서 사용자 정의 함수를 지정할 수 있도록 하고 있다.



[그림 5] HIVE 아키텍처 (출처 : Hive User Group Meeting ,2009)

3. IBM 분석 툴을 이용한 빅 데이터 활용 사례

2014년 월드컵과 2016년 올림픽을 준비하는 리오데자네이로는 지능형운영센터(IOC)를 통해 도시 관리와 긴급 대응 시스템을 갖췄다. 30여개에 이르는 여러 기관의 데이터와 프로세스를 지능형운영센터에 통합해 도시의 총체적인 움직임을 24시간 365일 모니터링할 수 있다.

◇ 리오데자네이로의 지능형운영센터(IOC) 전경 [사진=한국IBM]

[그림 6] 지능형운영센터(IOC)

IBM의 분석 솔루션이 적용된 지능형운영센터에는 교통, 전력, 홍수, 산사태 등의 자연재해와 수자원 등을 통합 관리할 수 있는 체계가 갖춰져 있다. IBM이 제공한 고해상도 날씨 예측 시스템과 수문학적 모델링 시스템은 날씨 및 수문 관련 방대한 데이터를 분석해 폭우를 48시간 이전에 예측한다. 강 유역의 지형측량 자료와 강수량 통계, 레이더 사진 등의 데이터에서 추출한 통합 수학적 모델에 기초해 강수량과 갑작스런 홍수를 예측한다. 뿐만 아니라 강수량과 교통체증, 정전 사태 등 도시에 영향을 미치는 상황들도 평가한다.

싱가폴에서는 차량의 기하급수적인 증가로 많은 교통체증을 겪고 있으며, 빅데이터는 이러한 싱가폴 행정기관에 새로운 해법을 제시하였다. 싱가포르는 빅데이터 분석을 통해 실시간 교통정보에서 한 단계 더 나아간 '교통량 예측 시스템(TPT)'을 운영하고 있다. 솔루션을 제공한 IBM에 따르면 전체적인 예측 결과는 85% 이상의 정확성을 보이고 있고 특히 교통량이 가장 많은 비즈니스 중심가에서는 정확도가 85% 이상으로 측정됐다.

미국에서 새로운 유통 트렌드를 발견하는데 이용된 빅데이터 분석 사례를 들어보자. 쇼핑과 세일을 대표하는 "블랙 프라이데이"는 연중 최대 규모라 알려져 있다. 하지만 IBM이 미국 전역의 500개 주요 유통기업에서 발생하는 하루 백만건 이상의 거래량과 테라바이트급(TB)의 빅데이터를 분석한 결과는 기존의 생각을 뒤집는 결과를 나타냈다. 즉, 추수감사절과 주말을 지낸 첫 월요일인 "사이버 먼데이"가 블랙 프라이데이 보다 주목받고 있음을 밝혀냈다.

4. 결론

현재 우리 사회는 데이터의 크기와 형태가 다양하고 데이터의 증가 속도가 가파른 이른바 '빅데이터 시대'에 놓여 있는 것이다. 오늘날 전 세계에서 다루어지는 디지털 정보량은 2년에 2배씩 증가하고 있다고 한다. 정보통신 기술이 다른 산업들과 융/복합되면서 방대한 량의 데이터들이 생성되고 있는 가운데 사회변화에 따른 삶의 질에 대한 욕구 및 현안 해결에 빅데이터들의 활용이 매우 증요한 과제로 떠오르고 있는 것이다.

빅데이터를 다루는 여러 플랫폼을 저장 시스템, 처리 프로세스(방식), 분석 메커니즘이라는 세 가지 측면으로 나누어 볼 수 있다. 본 연구에서는 위에서 기술된 3대 요소 중 처리 프로세스와 관련된 플랫폼 기술들에 비교, 분석하였다.

앞으로도 빅데이터 플랫폼에 대한 관심 및 활용성은 지속적으로 확대될 것이며 영역 또한 순수한 정보기술 영역을 넘어 모든 영역으로 광범위하게 적용될 것이다.

참 고 문 헌

[1] IDC보고서. (2012). 전세계 빅데이터 기술 및 서비스 전망, www.idc.com.

[2] 이성춘. (2012). 빅 데이터 활용과 통신산업에 대한 시사점.

[3] Gartner. (2011). Big Data Analytics. Gartner Group.

[4] IDG Korea. (2012). 빅데이터를 클라우드에서.

[5] Mckinsey Global Institute. (2011). Big Data: The next frontier for innovation, competition, and productivity.

[6] Warden, P. (2011). Big Data Glossary.

[7] www.itkorea.co.kr/ news.

[8] "Welcome to Apach Hadoop", http://hadoop.apache.org/

[9] Jeff kelly. (2012). Big Data: Hadoop, Business Analytics and Beyond.

이 성 훈

- 1998년 2월 : 고려대학교 컴퓨터학과 (이학박사)
- 1998년 3월 ~ 현재 : 백석대학교 정보통신학부 교수.
- 관심분야 : 분산 시스템, 무선 통신, 유전 정보, 웹서비스
- E-Mail : shlee@bu.ac.kr

이 동 우

- 2005년 2월 : 고려대학교 전산과학과 (이학박사)
- 1995년 3월 ~ 현재 : 우송대학교 컴퓨터정보학과 교수
- 관심분야 : 웹기반 분산시스템, 능동 시스템, 데이터베이스
- E-Mail : dwlee@wsu.ac.kr