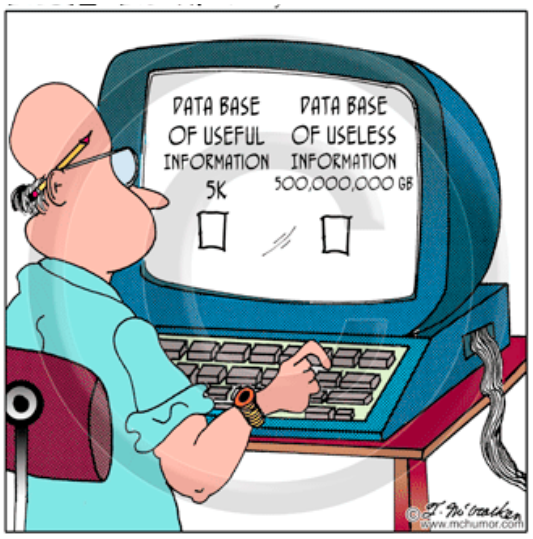
Big Data 시대의 기술

중앙연구소 Intelligent Knowledge Service
조성우

1. 시대의 화두 Big Data

최근 IT 분야의 화두가 무엇인지 물어본다면, 빅 데이터가 대답들 중 하나일 것이다. 20년 전의 PC의 메모리, 하드디스크의 용량과 최신 PC, 노트북 사양을 비교해보면 과거에 비해 데이터가 폭발적으로 늘어났다는 것을 실감할 수 있을 것이다. 특히 스마트 단말 및 소셜미디어 등으로 대표되는 다양한 정보 채널의 등장과 이로 인한 정보의 생산, 유통, 보유량의 증가는 계속적으로 데이터의 기하급수적인 증가를 이끌고 있다. 하지만 아래 그림과 같이 수 많은 데이터 중에서 가치 있는 데이터는 소수에 불과하다. 따라서 대용량 데이터를 처리하고, 의미 있는 데이터를 발굴하는 기술이 필요하다. 본고에서는 빅 데이터의 정의와 종류에 대해 알아보고, 거침없이 증가 중인 대용량의 데이터를 관리, 처리, 분석하기 위해 각광받고 있는 기술들을 살펴보려 한다.

[그림1] 실제 유용한 데이터는 소수에 불과하다



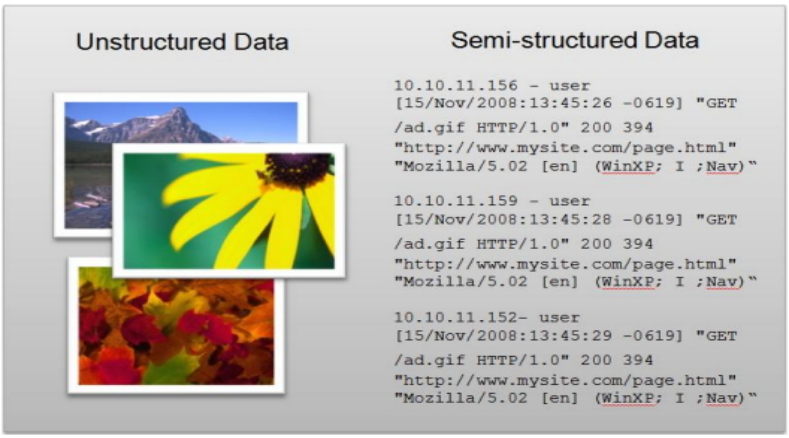
2. Big Data란 무엇인가?

빅 데이터는 어떻게 정의할 수 있을까? 사실 빅 데이터에 대해서 구체적이고 정량적인 정의가 합의된 바는 없다. 세계적인 컨설팅 기관인 McKinsey & Company는 2011년 5월에 발간한 보고서 'Big Data : The Next Frontier for Innovation, Competition, and Productivity'에서 "빅 데이터의 정의는 기존 데이터베이스 관리 도구의 데이터 수집, 저장, 관리, 분석하는 역량을 넘어서는 데이터셋Dataset 규모로, 그 정의는 주관적이며 앞으로도 계속 변화될 것이다. 또한 데이터량 기준에 대해 산업분야에 따라 상대적이며 현재 기준에서는 몇 십 테라바이트에서 수 페타바이트까지가 그 범위이다"라고 설명한다. 이렇게 데이터 단위의 관점에서 빅 데이터를 생각할 수도 있지만, 어떠한 종류의 데이터들이 빅 데이터를 구성하고 있는지 알아볼 필요도 있다.

빅 데이터의 데이터 종류는 정형화 정도에 따라 다음과 같이 분류된다.

정의	설명
정형 (Structured)	고정된 필드에 저장된 데이터. 관계형 데이터베이스 및 스프레드시트 등을 예로 들 수 있다.
반정형 (Semi-Structured)	고정된 필드에 저장되어 있지는 않지만, 메타데이터나 스키마 등을 포함하는 데이터. XML이나 HTML 텍스트 등을 예로 들 수 있다.
비정형 (Unstructured)	고정된 필드에 저장되어 있지 않은 데이터. 텍스트 분석이 가능한 텍스트 문서 및 이미지/동영상/음성 데이터 등을 예로 들 수 있다.

[그림2] 비정형 데이터와 반정형 데이터의 예

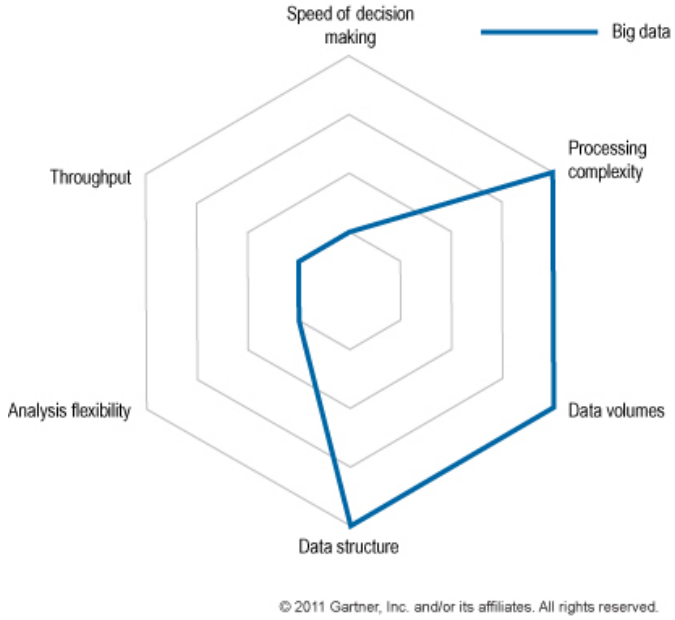


이중 비정형 데이터의 증가속도는 누구도 예측할 수 없는 정도이며, 비교적 선형적으로 증가하던 정형 데이터조차 연간 40~60%에 이르는 증가세를 보이고 있다.

그렇다면 빅 데이터를 처리는 기존 데이터 처리와 어떻게 다를까? IT 시장조사기관 Gartner는 2011년 1월 발간한 보고서 'Big Data Analytics'에서 기존 데이터 처리와 빅 데이터 처리에 대해 다음과 같은 차이점을 설명했다.

- 빠른 의사결정이 상대적으로 덜 요구된다 : 대용량 데이터에 기반한 분석 위주로, 장기적/전략적 접근이 필요하다. 따라서 기존의 데이터 처리에 요구되는 즉각적인 처리속도와는 달리, 즉각적인 의사결정이 상대적으로 덜 요구된다.
- 처리Processing 복잡도가 높다 : 다양한 데이터 소스, 복잡한 로직 처리, 대용량 데이터 처리 등으로 인해 처리 복잡도가 매우 높으며, 이를 해결하기 위해 통상적으로 분산 처리 기술이 필요하다.
- 처리할 데이터양이 방대하다 : 클릭스트림Clickstream 데이터를 예로 들면, 고객 정보수집 및 분석을 장기간에 걸쳐 수행해야 하므로 기존 방법과 비교해 처리해야 할 데이터양은 방대하다.
- 비정형 데이터의 비중이 높다 : 소셜 미디어 데이터, 로그 파일, 클릭스트림 데이터, 콜 센터 로그, 통신 CDR 로그 등 비정형 데이터 파일의 비중이 매우 높다. 처리의 복잡성을 증대시키는 요인이기도 하다.
- 처리/분석 유연성이 높다 : 잘 정의된 데이터 모델/상관관계/절차 등이 없어, 기존 데이터 처리방법에 비해 처리/분석의 유연성이 높은 편이다. 또한, 새롭고 다양한 처리방법의 수용을 위해, 유연성이 기본적으로 보장돼야 한다.
- 동시처리량Throughput이 낮다 : 대용량 및 복잡한 처리를 특징으로 하고 있어, 동시에 처리가 필요한 데이터양은 낮다. 따라서 (준)실시간 처리가 보장되어야 하는 데이터 분석에는 적합하지 않다.

[그림3] 빅 데이터의 성격을 한눈에 보여주는 그래프



3. Big Data를 위한 분석기법

위의 6가지 빅 데이터 처리의 특징을 만족시키기 위해 다양한 스토리지, 컴퓨팅 기술 및 분석기법들이 빅 데이터 처리를 위해 개발되었다. 빅 데이터 기술은 크게 분석기법과 분석 인프라 측면으로 나누어볼 수 있다.

대부분의 분석기법들은 통계학과 전산학, 특히 기계학습/데이터 마이닝 분야에서 이미 사용되던 기법들이며, 이 분석기법들의 알고리즘을 대규모 데이터 처리에 맞도록 개선하여 빅 데이터 처리에 적용시키고 있다. 최근 소셜미디어 등 비정형 데이터의 증가로 인해, 분석기법들 중에서 텍스트/오피니언 마이닝, 소셜 네트워크 분석, 군집분석 등이 주목을 받고 있다.

Text Mining

텍스트 마이닝은 비/반정형 텍스트 데이터에서 자연어처리Natural Language Processing 기술에 기반하여 유용한 정보를 추출, 가공하는 것을 목적으로 하는 기술이다. 텍스트 마이닝 기술을 통해 방대한 텍스트 뭉치에서 의미 있는 정보를 추출해 내고, 다른 정보와의 연계성을 파악하며, 텍스트가 가진 카테고리를 찾아내는 등, 단순한 정보 검색 그 이상의 결과를 얻어낼 수 있다. 컴퓨터가 인간이 사용하는 언어(자연어)를 분석하고 그 안에 숨겨진 정보를 발굴해 내기 위해 대용량

언어자원과 통계적, 규칙적 알고리즘이 사용되고 있다. 주요 응용분야로 문서 분류Document Classification, 문서 군집Document Clustering, 정보 추출Information Extraction, 문서 요약Document Summarization 등이 있다.

Opinion Mining

텍스트 마이닝의 관련 분야로는 오피니언 마이닝, 혹은 평판 분석Sentiment Analysis라고 불리는 기술이 있다. 오피니언 마이닝은 소셜미디어 등의 정형/비정형 텍스트의 긍정Positive, 부정Negative, 중립Neutral의 선호도를 판별하는 기술이다. 오피니언 마이닝은 특정 서비스 및 상품에 대한 시장규모 예측, 소비자의 반응, 입소문 분석Viral Analysis 등에 활용되고 있다. 정확한 오피니언 마이닝을 위해서는 전문가에 의한 선호도를 나타내는 표현/단어 자원의 축적이 필요하다.

Social Network Analytics

소셜 네트워크 분석은 수학의 그래프 이론Graph Theory에 뿌리를 두고 있다. 소셜 네트워크 연결구조 및 연결강도 등을 바탕으로 사용자의 명성 및 영향력을 측정하여, 소셜 네트워크 상에서 입소문의 중심이나 허브Hub 역할을 하는 사용자를 찾는데 주로 활용된다. 이렇게 소셜 네트워크 상에서 영향력이 있는 사용자를 인플루언서Influencer라고 부르는데, 인플루언서의 모니터링 및 관리는 마케팅 관점에서 중요하다고 할 수 있다.

Cluster Analysis

군집분석은 비슷한 특성을 가진 개체를 합쳐가면서 최종적으로 유사 특성의 군Group을 발굴하는데 사용된다. 예를 들어 트위터 상에서 주로 사진/카메라에 대해 이야기하는 사용자군이 있을 수 있고, 자동차에 대해 관심 있는 사용자군이 있을 수 있다. 이러한 관심사나 취미에 따른 사용자군을 군집분석을 통해 분류할 수 있다.

4. Big Data 분석 인프라 기술

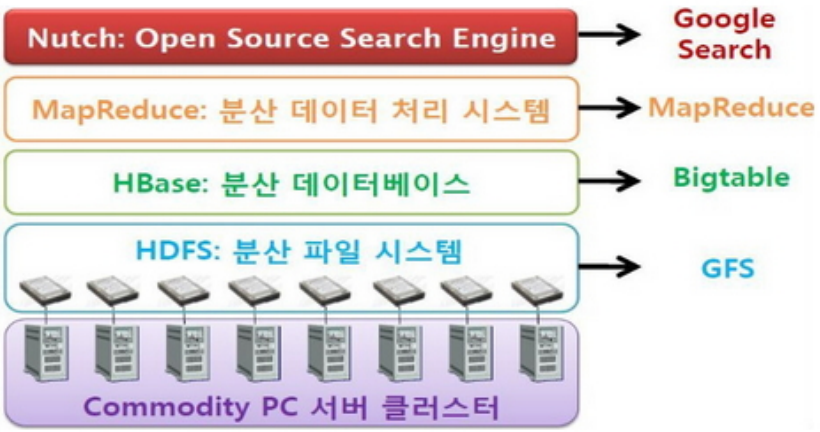
위의 분석기법들은 테라바이트 혹은 페타바이트 규모의 데이터에 적용되고 있다. 그렇다면 엄청난 규모의 빅 데이터 분석을 수행할 수 있는 인프라 기술은 어떤 것이 있을까? 일찍이 트래픽을 점유해왔던 회사들은 빅 데이터 처리를 위한

인프라 기술에 골몰해 왔다. 야후, 아마존, 구글 등의 회사들은 각자의 기술을 개발, 오픈소스화 하는데 앞장섰다.

Hadoop

하둡은 오픈소스Opensource 분산처리기술 프로젝트로, 현재 정형/비정형 빅 데이터 분석에 가장 선호되는 솔루션이라고 할 수 있다. 실제로 야후와 페이스북 등에 사용되고 있으며, 채택하는 회사가 늘어나고 있다. 주요 구성요소로 하둡 분산 파일 시스템인 HDFSHadoop Distributed File System, Hbase, MapReduce가 포함된다. HDFS와 Hbase는 각각 구글의 파일 시스템인 GFSGoogle File System와 빅 테이블Big Table의 영향을 받았다. 기본적으로 비용효율적인 x86 서버로 가상화된 대형 스토리지(HDFS)를 구성하고, HDFS에 저장된 거대한 데이터셋을 간편하게 분산처리 할 수 있는 Java 기반의 MapReduce 프레임워크를 제공한다. 이외의 Hadoop을 기반으로 한 다양한 오픈소스 분산처리 프로젝트가 존재한다.

[그림4] 하둡의 구조와 그에 대응하는 구글의 분산처리기술



R

오픈소스 프로젝트 R은 통계계산 및 시각화를 위한 언어 및 개발환경을 제공하며, R 언어와 개발환경을 통해 기본적인 통계기법부터 모델링, 최신 데이터 마이닝 기법까지 구현/개선이 가능하다. 이렇게 구현한 결과는 그래프 등으로 시각화할 수 있으며, Java나 C, Python 등의 다른 프로그래밍 언어와 연결도 용이하다. Mac OS, 리눅스/유닉스, 윈도우 등의 대부분의 컴퓨팅 환경을 지원하는 것도 장점이다. 위의 장점들로 인해 R은 통계분석 분야에서 인지도를 높여왔으며, 하둡 환경 상에서 분산처리를 지원하는 라이브러리 덕분에 구글, 페이스북, 아마존 등의 빅

데이터 분석이 필요한 기업에서 대용량 데이터 통계분석 및 데이터 마이닝을 위해 널리 사용되고 있다.

NoSQL

NoSQL은 Not-Only SQL, 혹은 No SQL을 의미하며, 전통적인 관계형 데이터베이스RDBMS와 다르게 설계된 비관계형 데이터베이스를 의미한다. 대표적인 NoSQL 솔루션으로는 Cassandra, Hbase, MongDB 등이 존재한다. NoSQL은 테이블 스키마Table Schema가 고정되지 않고, 테이블 간 조인Join 연산을 지원하지 않으며, 수평적 확장Horizontal Scalability이 용이하다는 특징을 가진다. 관계형 데이터베이스의 경우, 일관성Consistency(모든 노드는 같은 시간에 같은 데이터를 보여줘야 한다)과 유효성Availability(일부 노드가 다운되어도 다른 노드에 영향을 주지 않아야 한다)에 중점을 두고 있는 반면, NoSQL 기술은 분산가능성Partition Tolerance(네트워크 전송 중 일부 데이터를 손실하더라도 시스템은 정상 동작을 해야 한다)에 중점을 두고 일관성과 유효성은 보장하지 않는다. 이것은 일관성, 유효성, 분산가능성 중 2가지만 보장이 가능하다는 분산 데이터베이스 시스템 분야의 CAP 이론에 따른 것이다. 따라서 대규모의 유연한 데이터 처리를 위해서는 NoSQL 기술이 적합하지만, 안정성이 중요한 시스템에서는 오랫동안 검증된 관계형 데이터베이스를 채택할 필요가 있다.

4. Big Data 시대를 맞이하여

본고에서는 빅 데이터 시대를 맞이하기 위해 필요한 분석기술과 인프라 기술에 대해서 살펴 보았다. 물론 본문에서 빅 데이터 처리를 위한 모든 기술에 대해 논하지 못했지만, 현재 대표적으로 알아야 할 기술에 대해선 어느 정도 논의했다고 생각된다. 이런 다양한 기술 및 기법들을 실제 빅 데이터 분석에 활용하기 위해서는 보유하고 있는 데이터의 성격과 기술의 장단점을 잘 파악하고 적용하는 것이 중요할 것으로 보인다. 또한 기술 및 인프라 관점에서만 빅 데이터를 바라보는 것이 아니라 빅 데이터에 기반한 새로운 서비스에 대해서도 고민이 필요하다. 고도화된 빅 데이터 처리 기술을 기반으로 한 새로운 서비스들이 앞으로 열어갈 미래가 기대된다.

<참고문헌 및 웹사이트>

[1] 하둡 기술 연계한 데이터 분석, 김희배, 2011년 9월
[2] Big Data Analytics, Gartner, 2011년 1월
[3] Big Data : The Next Frontier for Innovation, Competition, and Productivity, McKinsey & Company, 2011년 5월
[4] Managing Big Data with Hadoop & Vertica, Vertica Systems, 2009년 10월
[5] http://hadoop.apache.org/
[6] http://www.r-project.org/
[7] http://cassandra.apache.org/
[8] http://www.mongodb.org/