

실무에 적용 가능한 Big Data 분석 개론

빅데이터 기술



한국기술교육대학교
온라인평생교육원

■ 빅데이터 기술 분류와 빅데이터 생성 및 수집기술

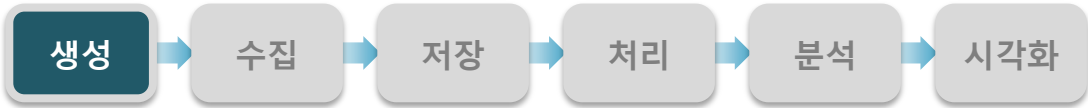
1. 빅데이터 기술 분류

과정	설명	해당기술
생성	조직의 내부와 외부에 존재하는 여러 데이터를 생성하는 기술	<ul style="list-style-type: none"> • 데이터베이스(Database) • 파일관리시스템(File Management system) • 인터넷으로 연결된 파일 등
수집	조직의 내부와 외부에서 생성되는 여러 데이터 소스로부터 필요로 하는 데이터를 검색하여 수동 또는 자동으로 수집하는 과정과 관련된 기술로 단순 데이터 확보가 아닌 검색, 수집, 변환을 통해 정제된 데이터를 확보하는 기술	<ul style="list-style-type: none"> • 로그 수집기 • 크롤링 • 센싱 • RSS Reader, Open API • ETL(Extraction, Transformation, Loading) 등
저장	작은 데이터라도 모두 저장하고 실시간으로 저렴하게 데이터를 처리하고 처리된 데이터를 더 빠르고 쉽게 분석하도록 효율적으로 저장하는 기술	<ul style="list-style-type: none"> • 분산 파일 시스템(Distributed File System) • NoSQL • 병렬 DBMS 등
처리	엄청난 양의 데이터의 저장, 수집, 관리, 유통, 분석을 처리하는 일련의 기술	<ul style="list-style-type: none"> • 실시간 처리 • 분산병렬처리 • 맵리듀스(MapReduce) 등
분석	데이터를 효율적으로 정확하게 분석하여 비즈니스 등의 영역에 적용하기 위한 기술로 이미 여러 영역에서 활용해온 기술	<ul style="list-style-type: none"> • 통계분석 • 데이터 마이닝 • 텍스트 마이닝 • 평판분석 • 소셜 네트워크 분석 등
시각화	자료를 시각적으로 묘사하는 기술로, 빅데이터는 기존의 단순 선형적 구조의 방식으로 표현하기 힘들기 때문에 필수적	<ul style="list-style-type: none"> • 정보 편집 기술 • 정보 시각화 기술 • 시각화 도구 등

<한국정보화진흥원, 빅데이터 기술 분류 및 현황, 2013>

■ 빅데이터 기술 분류와 빅데이터 생성 및 수집기술

2. 빅데이터 생성 및 수집 기술

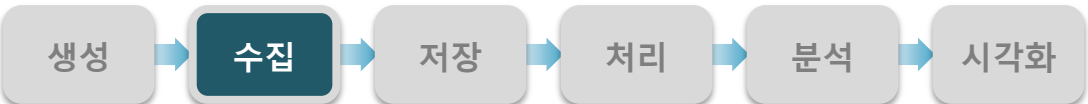


- 폭발적으로 증가하고 있는 데이터들을 존재하게 하는 단계
- 조직의 내부와 외부에 존재하는 **수많은 데이터를 수집하는 단계**

해당 기술

공유되어 사용될 목적으로
통합하여 관리되는 데이터의
집합

- 내부 데이터(정형 데이터)
 - 데이터베이스(Database) 관리시스템
 - 파일관리시스템(File Management system)
- 외부 데이터(비정형 데이터)
 - 인터넷으로 연결된 파일 등



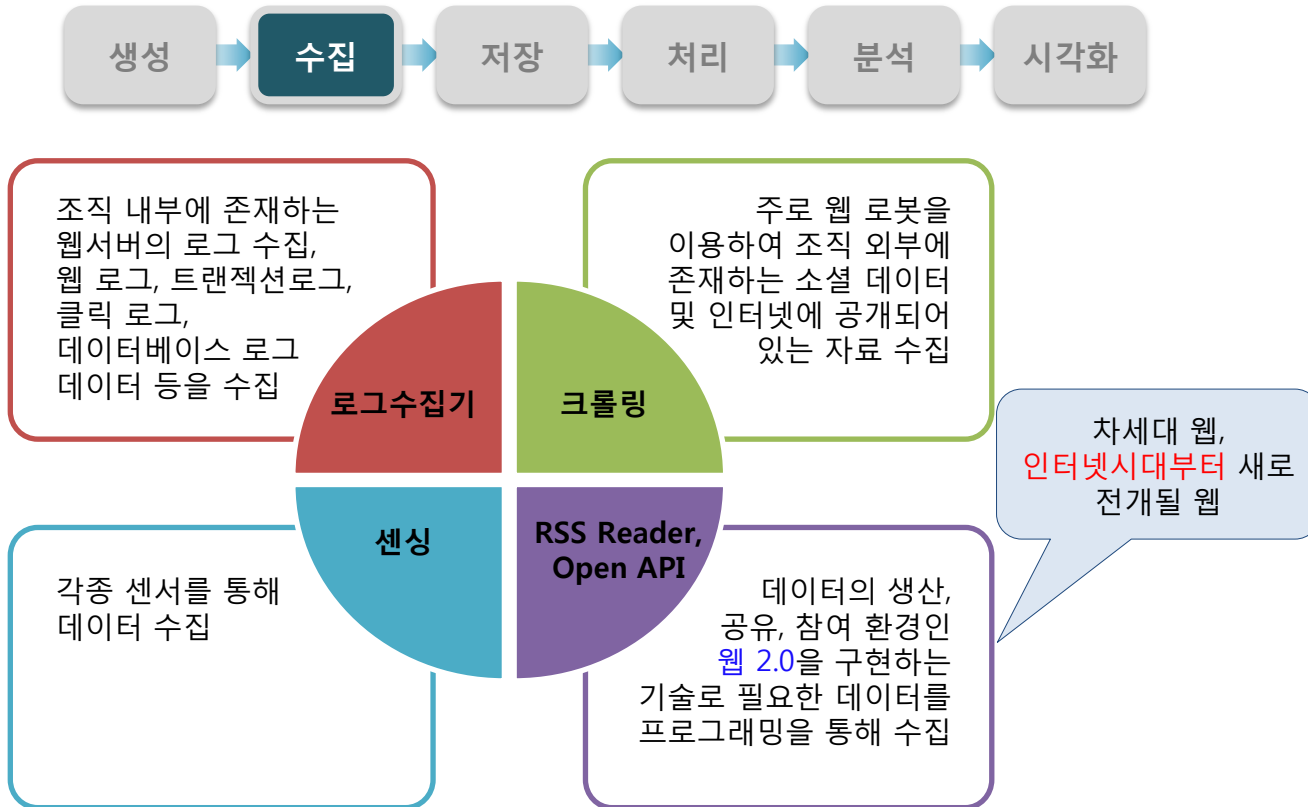
- 조직의 내부와 외부에서 생성되는 여러 데이터 소스로부터 필요로 하는 데이터를 검색하여 **수동 또는 자동으로 수집하는 과정과 관련된 기술**
- 단순 데이터 확보가 아닌 검색, 수집, 변환을 통해 **정제된 데이터를 확보**하는 기술을 의미함

해당 기술

- 내부 데이터(정형 데이터)
 - 로그 수집기
- 외부 데이터(비정형 데이터)
 - 크롤링
 - 센싱
 - RSS Reader, Open API
 - ETL(Extraction, Transformation, Loading) 등

■ 빅데이터 기술 분류와 빅데이터 생성 및 수집기술

2. 빅데이터 생성 및 수집 기술



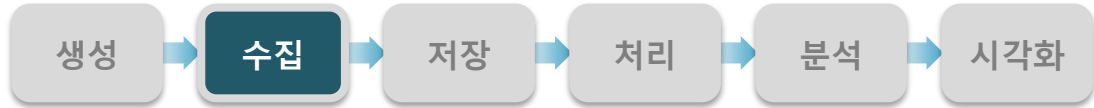
<한국정보화진흥원, 빅데이터 기술 분류 및 현황, 2013>

▶ RSS Reader

- **RDF Site Summary**(RDF 사이트 요약)
 - **Rich Site Summary**(풍부한 사이트 요약)
 - **Really Simple Syndication**(초간편 배급)
-
- 사이트에서 제공하는 주소를 RSS Reader 프로그램에 등록하면, PC나 휴대폰 등을 통하여 자동으로 전송된 콘텐츠를 이용할 수 있음
 - RSS Reader에 제목과 내용 요약, 날짜 등 배포에 필요한 최소한의 정보가 이메일의 목록처럼 나열되고, 사용자가 원하는 콘텐츠를 클릭하여 해당 페이지로 접속하는 방식임

■ 빅데이터 기술 분류와 빅데이터 생성 및 수집기술

2. 빅데이터 생성 및 수집 기술



처리 과정이 원활하게 진행되려면 시스템이 원하는 처리 속도에 맞게 데이터가 수집되어야 함

데이터 수집 기술



빅데이터 처리 과정의 시작

실시간성이 요구되는 기술들이
실시간 분석 서비스들과 함께 부각되고 있음

실무에 적용 가능한 Big Data 분석 개론

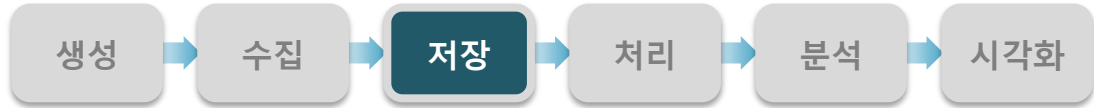
빅데이터 기술



한국기술교육대학교
온라인평생교육원

■ 빅데이터 저장 기술과 처리 기술

1. 빅데이터 저장 기술



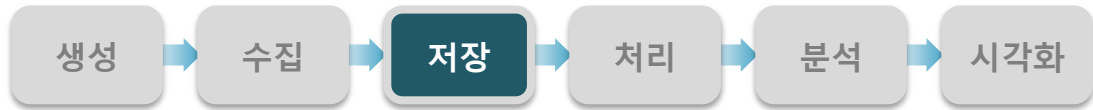
- 작은 데이터라도 모두 저장하고 실시간으로 저렴하게 데이터를 처리하고 처리된 데이터를 더 빠르고 쉽게 분석하도록 효율적으로 저장하는 기술을 의미함
- 빅데이터의 대용량, 비정형, 실시간성의 속성을 수용할 수 있는 저장방식이 필요함

해당 기술

- 분산 파일 시스템(Distributed File System)
 - 하둡 분산 파일 시스템 (HDFS : Hadoop Distributed File System)
 - 구글 파일 시스템(GFS : Google File System) 등
- NoSQL
 - H베이스(Hbase), 카산드라(Cassandra), 몽고DB(MongoDB) 등
- 병렬 DBMS
 - 버티카(Vertica), 그린플럼(Greenplum), 넷티자(Netezza) 등

■ 빅데이터 저장 기술과 처리 기술

1. 빅데이터 저장 기술



분산 파일 시스템(Distributed File System)

컴퓨터 네트워크로 공유하는
여러 호스트 컴퓨터 파일에 접근할 수 있는 파일 시스템



하둡 분산 파일 시스템 (HDFS : Hadoop Distributed File System)

- 하둡(Hadoop)
 - 대량의 자료를 저장하고 처리할 수 있는 큰 컴퓨터 클러스터에서 동작하는 분산 응용 프로그램을 지원하는 **자바 소프트웨어 프레임 워크**
 - 7년 간 개발되면서 **개방형 프레임 워크**로 빅데이터 시대를 이끌고 있음
 - 하둡을 중심으로 한 새로운 제품군들이 등장하고 있음
- 하둡 분산 파일 시스템(HDFS : Hadoop Distributed File System)
 - 이기종간의 하드웨어로 구성된 컴퓨터 클러스터에서 **대용량 데이터 처리**를 위하여 개발된 분산 파일 시스템



구글 파일 시스템(GFS : Google File System)

- **구글(Google)**
 - 웹 검색, 클라우드 컴퓨팅, 광고를 주 사업 영역으로 하는 미국의 다국적 회사
- **구글 파일 시스템(GFS : Google File System)**
 - 구글에 의해 자기 회사 사용 목적으로 개발된 분산 파일 시스템
 - 일반 상용 하드웨어를 이용하여 대량의 서버를 연결했기 때문에 데이터에 대한 **접근이 효율적이고 안정적임**

■ 빅데이터 저장 기술과 처리 기술

1. 빅데이터 저장 기술



Not only SQL(SQL뿐만 아니라)
SQL은 표준으로 채택하고 있는 특수 목적의 프로그래밍 언어임
➡ NoSQL은 기존과는 다른 새로운 데이터 저장 기술

- ✓ 기존 문제를 개선, 보완하기 위해서 사용된 새로운 데이터 저장기술
- ✓ 비관계형 데이터베이스를 지칭하는 분산 환경의 데이터 저장소
- ✓ NoSQL에서 데이터 저장
 - 다수의 서버에 분산해서 저장하여 속도가 빠름
 - 트랜잭션이 클러스터를 구성하는 전체 서버에 분산되기 때문에 다수의 클라이언트가 동시에 접속해서 사용하여도 됨

Not only SQL, 기존과는 다르다!
(SQL ➡ 기존 프로그래밍 언어)

APACHE HBASE H베이스(Hbase)

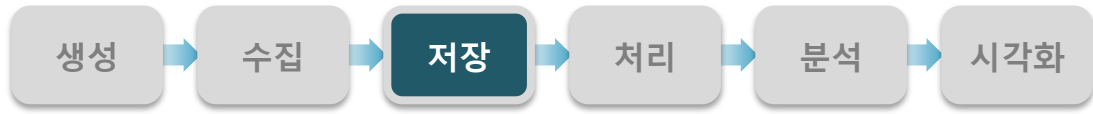
- 하둡 분산 파일 시스템(HDFS : Hadoop Distributed File System)에서 동작되는 오픈 소스, 분산, 비관계형 데이터베이스

무상으로 공개된 소스 코드

- 구글의 '빅데이터'를 참고로 개발됨
- 파워셋에서 개발했으며, 현재는 아파치 소프트웨어 재단에서 한 프로젝트로 관리함
- 데이터를 많이 사용하는 웹사이트에서 사용됨

■ 빅데이터 저장 기술과 처리 기술

1. 빅데이터 저장 기술



NoSQL

Not only SQL, 기존과는 다르다!
(SQL → 기존 프로그래밍 언어)



Cassandra 카산드라(Cassandra)

- 분산 시스템에서 대용량 데이터를 처리할 수 있도록 설계된 오픈 소스 데이터베이스 관리 시스템
 - 아마존의 다이나모(Dynamo)와 구글의 빅테이블(BigTable)의 장점만을 수용하여 발전한 형태인 NoSQL의 대표적인 데이터베이스
 - 원래 페이스북에서 개발했으며 지금은 아파치 소프트웨어 재단에서 한 프로젝트로 관리함

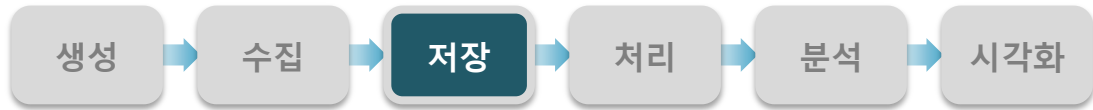


mongoDB 몽고DB(MongoDB)

- 신뢰성과 확장성에 기반한 문서 지향 데이터베이스
 - 방대한 양의 데이터에서 낮은 관리 비용과 사용 편의성을 목표로 함
- 가장 유명한 NoSQL 데이터베이스 시스템
 - 10gen이 오픈 소스로 개발한 것으로 상업적인 지원이 가능함

■ 빅데이터 저장 기술과 처리 기술

1. 빅데이터 저장 기술



병렬 DBMS

다수의 마이크로프로세서를 사용하여 여러 디스크의 질의, 갱신, 입출력 등 데이터베이스 처리를 동시에 수행하는 데이터베이스 시스템



버티카(Vertica)

- 휴렛패커드 (HP : Hewlett-Packard Company) 자회사
- 빠른 분석을 위한 컬럼 기반의 대규모 병렬처리용 데이터베이스



Greenplum

그린플럼(Greenplum)

- EMC 코퍼레이션 자회사
- 관계형 데이터베이스와 하둡을 한 장비에 넣음
 - 데이터웨어하우징 (DW)업계 최초



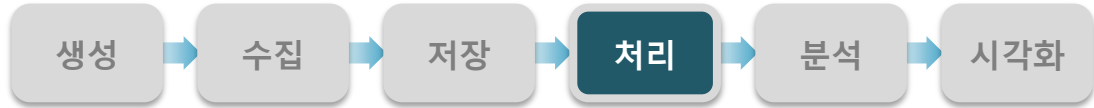
NETEZZA[®]
an IBM[®] Company

네티자(Netezza)

- IBM 자회사
- 기존 데이터베이스를 이용하는 타사와 달리 데이터웨어하우징의 처리 고속화를 위해서 설계된 제품을 제공함

■ 빅데이터 저장 기술과 처리 기술

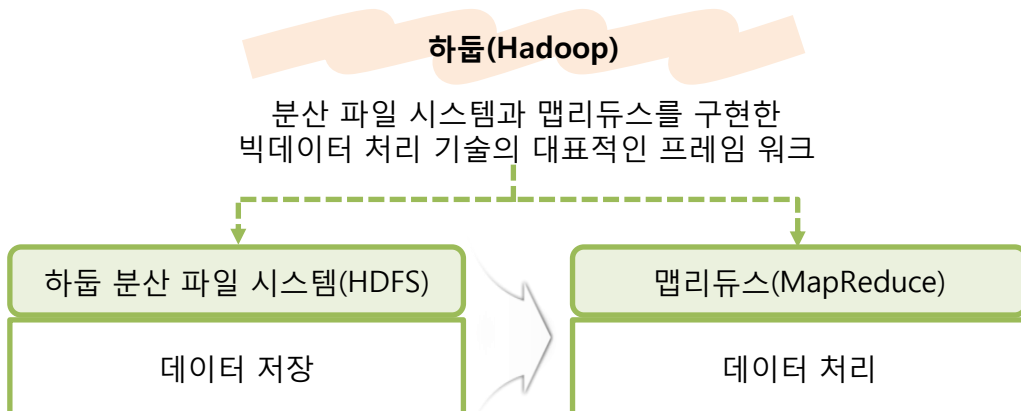
2. 빅데이터 처리 기술



- 빅데이터에서 유용한 정보 및 숨어있는 지식을 찾아내기 위한 **데이터 가공 및 분석과정을 지원**하는 기술을 의미함
- 대규모 데이터 처리를 위한 확장성, 데이터 생성 및 처리 속도를 해결하기 위한 처리시간 단축 및 실시간 처리 지원, 비정형 데이터 처리 지원 등이 필요함

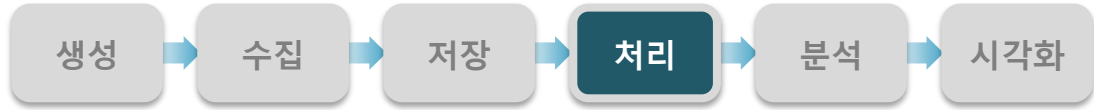
해당 기술

- 하둡(Hadoop)
- NoSQL
- 구글 맵리듀스(MapReduce) 등



■ 빅데이터 저장 기술과 처리 기술

2. 빅데이터 처리 기술



하둡 분산 파일 시스템(HDFS)

✓ 파일을 블록단위로 나누어 각 노드 클러스터에 저장함

✓ 데이터 유실을 막고 부하처리를 위해 각 블록의 복사본을 생성함

적은 비용으로
빅데이터의 처리가
가능함

높은 사용 편의성을
제공함

유실이나 장애
발생 시 복구가
가능함

맵리듀스(MapReduce)

구글이 발표한 대표적인 빅데이터 병렬처리 모델

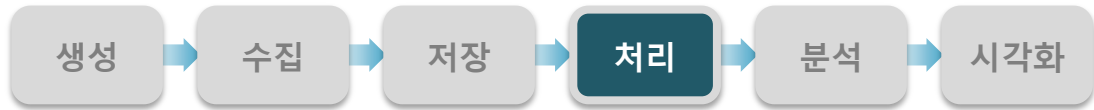
✓ 대용량 데이터를 빠르고 안전하게 처리하기 위한 분산 프로그래밍 모델임

✓ 맵(Map)함수와 리듀스(Reduce)함수 기반으로 구성되는 데이터를 병렬 처리하는 기술임

✓ 하둡에서도 구현됨

■ 빅데이터 저장 기술과 처리 기술

2. 빅데이터 처리 기술



NoSQL

Not only SQL

기존과 다르게 설계된 비관계형 데이터베이스

APACHE
HBASE

 **Cassandra**

 **mongoDB**

빅데이터 처리 기술의 동향

하둡을 포함한 오픈소스 진영을 중심으로 다양한 기술이 빠르게 진화하고 있음

아직까지 상용 솔루션보다 오픈소스의 비중이 높음

실무에 적용 가능한 Big Data 분석 개론

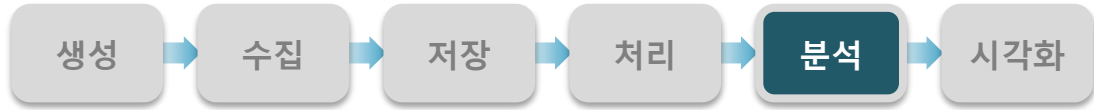
빅데이터 기술



한국기술교육대학교
온라인평생교육원

▣ 빅데이터 분석 기술과 시각화 기술

1. 빅데이터 분석 기술



- 데이터를 효율적으로 정확하게 분석하여 비즈니스 등의 영역에 적용하기 위한 기술로 이미 여러 영역에서 활용해온 기술
- 빅데이터로부터 숨어있는 패턴과 지식을 찾아내기 위한 기술을 의미함
- 찾아진 패턴과 지식을 토대로 비즈니스 영역에서는 의사결정을 수행함

해당 기술

- 통계분석
- 데이터 마이닝
- 텍스트 마이닝
- 평판분석
- 소셜 네트워크 분석 등

통계분석



다양한 분석에서 활용되는 기술



통계적 컴퓨팅에 사용되는 R, SAS 등을 통하여 다양한 통계기법으로 분석할 수 있음

R

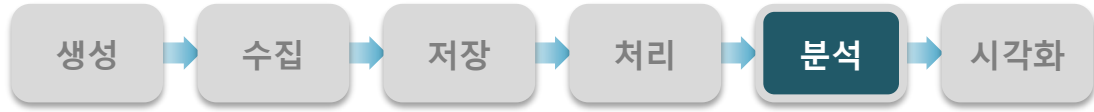
- 빅데이터 분석 기술 도구
- 통계계산 및 시각화를 위한 언어 및 개발 환경을 제공할 경우
 - 기본적인 통계 기법부터 데이터 마이닝 기법까지 구현이 가능함

SAS (Statistical Analysis System)

- 미국 노스캐롤라이나 주립 대학교에서 1967년 원형이 개발되고 SAS Institute사에서 기능을 확장함
- 통계 해석을 중심으로 한 소프트웨어 패키지

▣ 빅데이터 분석 기술과 시각화 기술

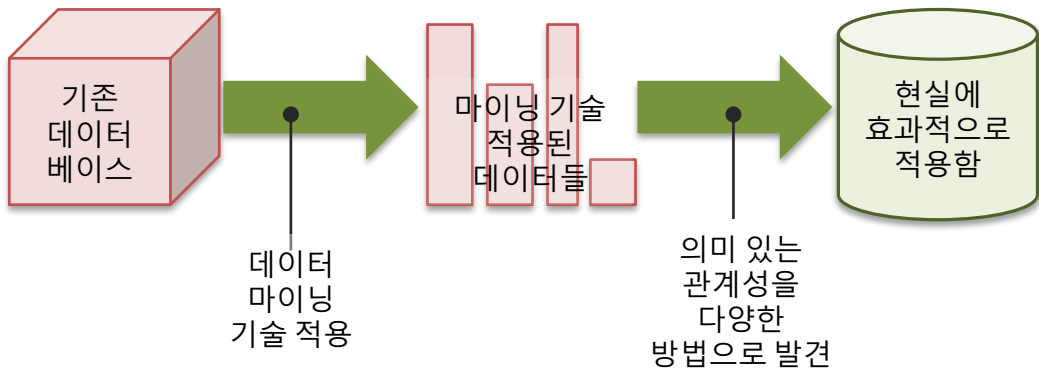
1. 빅데이터 분석 기술



데이터 마이닝

✓ 통계 및 수학적 기술뿐 아니라 기계학습, 패턴인식, 신경망 등의 기술들을 이용함

✓ 대용량의 데이터에 숨겨진 의미 있는 패턴, 추세, 지식들을 발견함



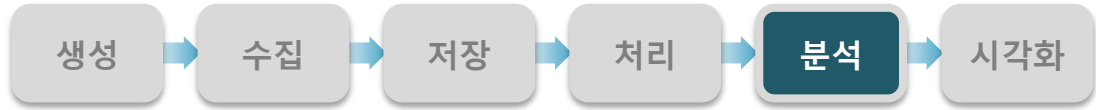
텍스트 마이닝



- 구조화되지 않은 대규모의 텍스트 집합으로부터 새로운 지식을 발견하는 기술
- 텍스트 문서로부터 정보 검색, 정보 추출, 체계화 및 분석을 포함함

▣ 빅데이터 분석 기술과 시각화 기술

1. 빅데이터 분석 기술



평판분석



- 소셜 미디어 등의 정형 또는 비정형 텍스트의 긍정, 부정, 중립의 선호도를 판별하는 분석 기술
- 주로 특성 서비스 및 상품에 대한 시장 규모 예측, 소비자의 반응, 입소문 분석 등에 활용됨

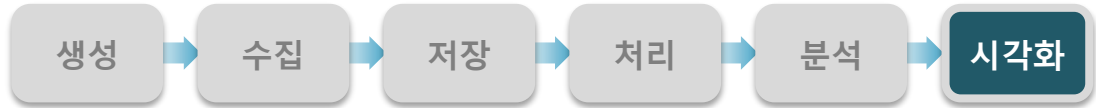
소셜 네트워크 분석





- 소셜 네트워크 연결 구조 및 연결 강도 등을 바탕으로 사용자의 명성 및 영향력을 분석하는 기술
- 주로 마케팅을 위하여 소셜 네트워크 상에서 입소문의 중심이나 허브 역할을 하는 사용자를 찾는데 활용됨

▣ 빅데이터 분석 기술과 시각화 기술

2. 빅데이터 시각화 기술



- 데이터 분석결과를 쉽게 이해할 수 있도록 시각적인 수단으로 정보를 전달하는 과정
- 데이터 안의 수많은 패턴들을 시각화하여 핵심개념과 아이디어를 직관적이고 명확하게 이해할 수 있는 기술을 의미함

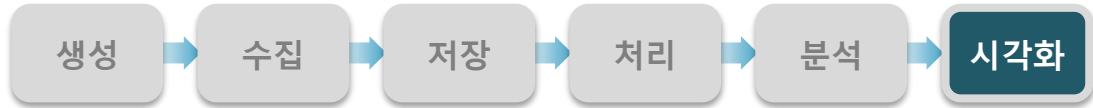
 가장 중요한 기술 분야 

해당 기술

- 정보 편집 기술
- 정보 시각화 기술
- 시각화 도구 등

■ 빅데이터 분석 기술과 시각화 기술

2. 빅데이터 시각화 기술



정보 편집 기술

시각적 매핑, 스토리 텔링 등의 방법이 활용됨

시각적 매핑

- 시간에 따른 데이터 변화를 표현하는 방법

스토리 텔링

- Story + Telling, 이야기하듯의 의미
- 상대방에게 알리고자 하는 바를 재미있고 생생한 이야기로 설득력 있게 전달하듯이 구성하는 것

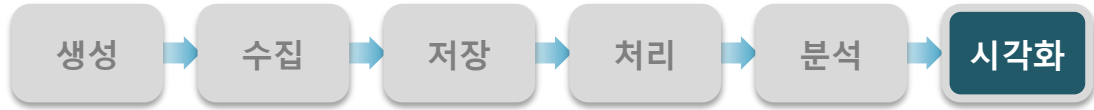


시각적 매핑의 예 - 하루일과

<한국정보화진흥원, 빅데이터 기술 분류 및 현황, 2013>

▣ 빅데이터 분석 기술과 시각화 기술

2. 빅데이터 시각화 기술



정보시각화 기술

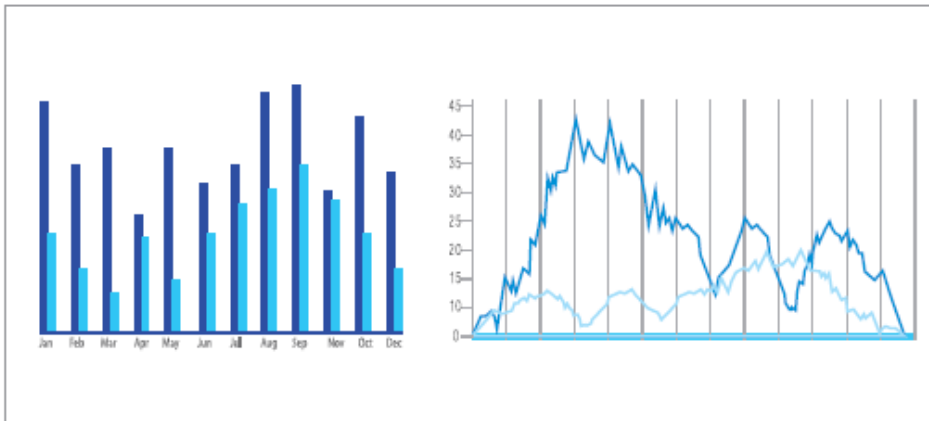
시간 시각화, 분포 시각화 등의 방법이 활용됨

시간 시각화

- 시간에 따른 데이터 변화를 표현하는 방법

분포 시각화

- 최대, 최소, 전체분포로 분류하여 시각적으로 표현하는 방법

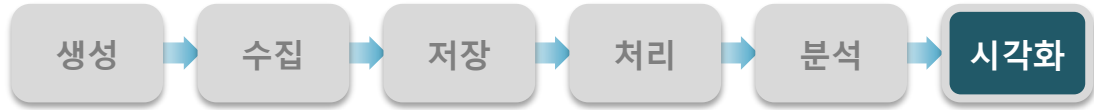


누적막대그래프와 시계열 그래프 예

<한국정보화진흥원, 빅데이터 기술 분류 및 현황, 2013>

▣ 빅데이터 분석 기술과 시각화 기술

2. 빅데이터 시각화 기술



시각화 도구

마이크로소프트 엑셀, 구글 스프레드시트, IBM의 매니아이즈 등의 소프트웨어 도구가 시각화에 사용됨



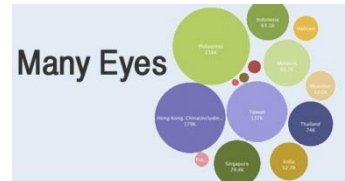
마이크로소프트 엑셀

- 윈도 환경의 스프레드시트 프로그램



구글 스프레드시트

- 연산 및 표를 작성하고 그래프를 그리는 소프트웨어



IBM의 매니아이즈

- 다양한 범위의 시각 자료를 보고 쉽게 편집할 수도 있는 사이트