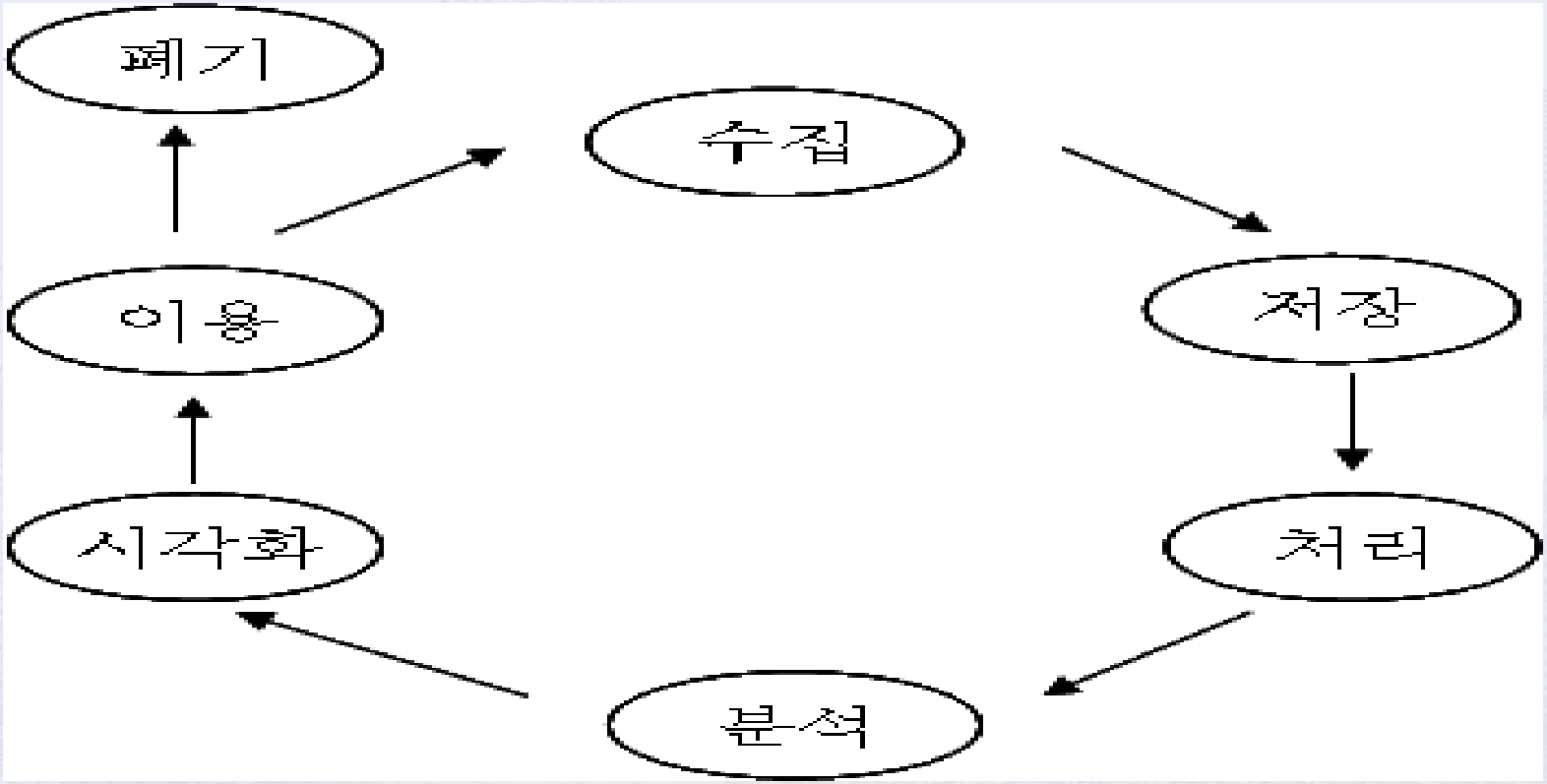
빅데이터

1. 빅데이터 분석 프로세스의 개념

- 빅데이터 분석의 주요 목적은 데이터 분석 전문가가 기존의 전통적인 비즈니스 인텔리전스(BI: business intelligence) 프로그램이 시도하지 않았던 웹 서버 로그, 인터넷 클릭 정보, 소셜 미디어 활동 보고서, 이동 전화 통화 기록, 또는 센서들이 감지한 정보 등의 새로운 종류의 데이터(data sources)나 많은 양의 트랜잭션 데이터(transaction data)를 분석할 수 있도록 하여
- 기업이 경영과 관련하여 더 좋은 의사결정을 하도록 도와주는 것이다.

1. 빅데이터 분석 프로세스의 개념

빅데이터 처리의 순환과정

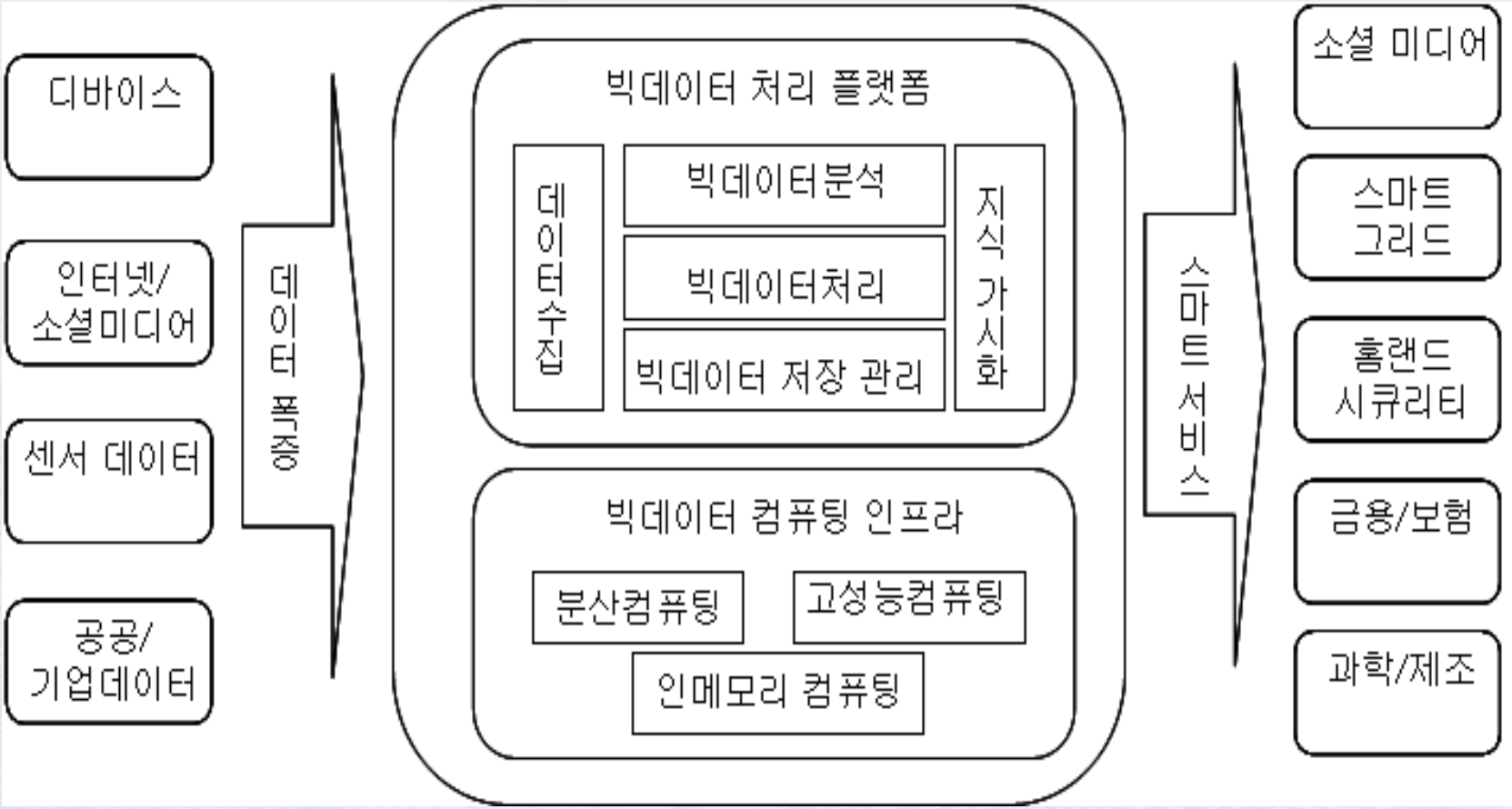


2. 빅데이터 플랫폼

- 빅데이터 처리 플랫폼은 빅데이터 수집, 빅데이터 저장/관리 기술, 빅데이터 처리 기술, 빅데이터 분석 기술 및 지식 시각화 기술 등을 적용하여 구현한다.
- 빅데이터로부터 지식을 얻어 활용하기까지는 여러 단계가 필요하고, 그 단계마다 수많은 기술이 활용된다.
- 빅데이터 플랫폼은 데이터를 수집해서 지식을 발굴하는 데 필요한 빅데이터 처리 플랫폼 기술, 대용량의 고속 저장 공간 및 고성능의 계산 능력을 갖춘 컴퓨터 , 컴퓨팅 기반을 제공하는 빅데이터 컴퓨팅 인프라 기술로 구성된다.

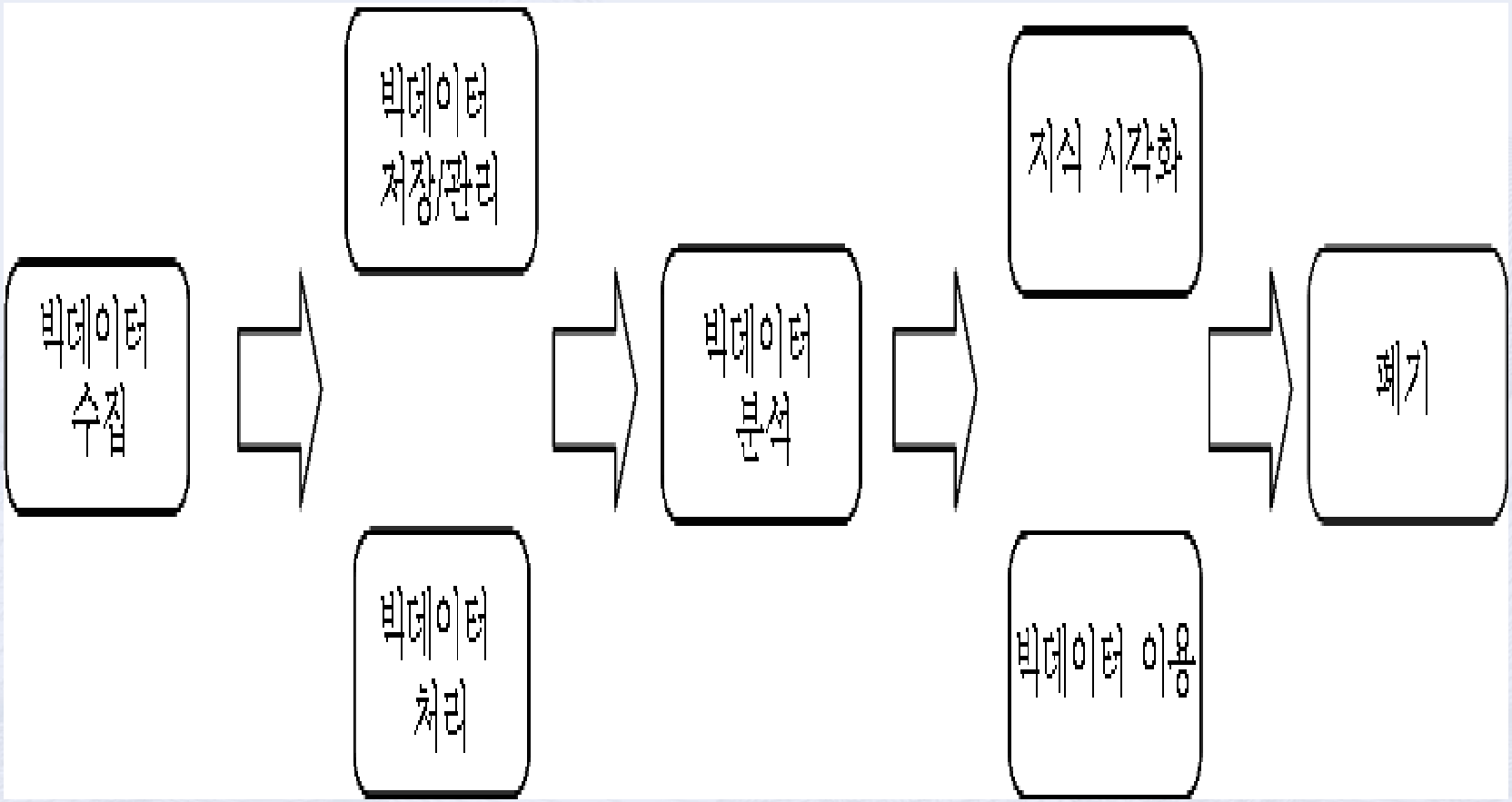
2. 빅데이터 플랫폼

빅데이터 플랫폼 개념도



3. 빅데이터 분석 프로세스 절차

빅데이터 분석 처리 프로세스



3. 빅데이터 분석 프로세스 절차

➢ 3.1 빅데이터 수집 (Big Data Collection)

- 일반적으로 말하는 데이터는 기업이나 조직 내부에 있는 정보시스템에 저장된 정형화된 데이터로서 데이터 수집에 큰 노력을 기울이지 않아도 수집이 가능하고 수집하고자 하는 데이터의 형식도 개발 단계에서부터 향후에 분석하기에 적합한 형식에 갖춘 정형화된 형식의 로그로 구현하기 때문에 수집 후에 데이터를 가공하는 데에 큰 노력이 들어가지 않는다.

- 하지만 빅데이터는 내부 조직에 있는 정형화된 데이터뿐만 아니라, 조직 외부에 존재하는 무한한 데이터 중에서 조직이 필요로 하는 데이터를 발견하여, 이를 수집하고 수집된 정보를 분석을 위한 특정 데이터 형식으로 변환하는 과정을 거쳐야 한다. 따라서 빅데이터 수집이란 단순히 데이터를 확보하는 기술이 아니라 데이터를 검색하여 수집하고 변환 과정을 통해 정제된 데이터를 확보하는 과정을 말한다.

3. 빅데이터 분석 프로세스 절차

빅데이터 수집의 세부 절차

3. 빅데이터 분석 프로세스 절차

- **3.1.1) 수집 대상 데이터 선정**
  - 빅데이터 수집은 빅데이터 분석이나 서비스를 제공할 때에 서비스의 품질을 결정하는 중요한 핵심 단계로 수집 대상 분야에 분석 경험이 많은 전문가의 의견을 반영하여 분석 목적에 맞는 데이터를 선정하여야 한다.
  - 수집 대상을 선정할 때는 대상 데이터가 수집이 가능하고 사용이 가능한지의 여부, 이용 목적에 맞는 세부 항목이 포함되어 있는지 여부 그리고 개인 정보 침해의 여부나 수집 비용을 고려해야 한다.

3. 빅데이터 분석 프로세스 절차

- **3.1.2) 수집 세부 계획 수립**
  - 이 단계에서는 데이터 소유자를 확인하고 대상 데이터가 내부 데이터인지 외부 데이터인지 또는 수집 대상 데이터의 유형과 데이터 포맷을 확인하여 적정한 수집 기술을 선정하여야 한다.

3. 빅데이터 분석 프로세스 절차

데이터 유형에 따른 수집 기술

데이터 유형	데이터 종류	수집기술
정형 데이터	RDB. 스프레드 시트	ETL, FTP, Open API
반정형 데이터	HTML. XML. JSON. 웹문서. 웹로그. 센서 데이터	Crawling, RSS, Open API, FTP
비정형 데이터	소셜 데이터. 문서(워드, 한글). 이미지. 오디오. 비디오. IoT	Crawling, RSS, Open API, Streaming, FTP

3. 빅데이터 분석 프로세스 절차

3. 1. 3) 데이터 수집 실행

(1) 능동적 데이터 수집과 수동적 데이터 수집

- 데이터 수집 실행은 위에서 언급한 다양한 기술을 적용하여 수행을 하게 되는데 데이터를 수집하는 주체의 능동성 여부에 따라서 능동적 데이터 수집과 수동적 데이터 수집으로 분류할 수 있다.

(2) 내부 또는 외부 데이터 수집

- 데이터 수집은 데이터 소스의 위치에 따라 내부 데이터 수집과 외부 데이터 수집으로 구분할 수 있다. 내부 데이터 수집은 주로 자체적으로 보유한 내부 파일 시스템이나 데이터베이스 관리 시스템, 센서 등에 접근하여 데이터를 수집하는 것을 의미하고, 외부 데이터 수집은 인터넷으로 연결된 외부에서 데이터를 수집하는 것을 의미한다.

3. 빅데이터 분석 프로세스 절차

3.1. 4) 빅데이터 변환/통합

- 빅데이터의 변환은 데이터를 수집하는 과정에서 컴퓨터가 바로 처리할 수 없는 **비정형 데이터를 구조적 형태로 전환**하여 저장하는 것을 말한다.
- 또한 빅데이터 변환은 빅데이터 **정제**(cleansing)를 포함한다. 이것은 비정형 데이터 (unstructured data)를 정제하거나 또는 정형적 데이터(structured data)에서 측정값이 빠져 있다거나, 형식이 다르다거나, 내용 자체가 틀린 데이터를 고쳐주는 과정을 말한다.
- 데이터의 통합은 빅데이터를 효과적으로 분석하기 위하여 레거시 데이터 간 **통합**을 말한다.

3. 빅데이터 분석 프로세스 절차

빅데이터 수집을 위한 변환 및 통합

ETL (Extraction, Transformation, Load)	메인 프레임, ERP, CRM, Flat file, Excel 파일 등으로부터 데이터를 추출하여 목표하는 저장소의 데이터 형태로 변형한 후 목표 저장소(DW)에 저장
비정형 -> 정형	비정형 데이터는 (비구조적 데이터 저장소에 저장하거나)어느 정도 구조적인 형태로 변형하여 저장 ex) Scribe, Flume, chuckwa 등 오픈 소스 솔루션
레거시 데이터와 비정형 데이터간의 통합	데이터를 분석하기 위해서는 수집된 정형의 레거시 데이터와 비정형 데이터간의 통합 예) Sqoop : RDBMS와 HDFS간의 데이터를 연결해 주는 기능으로 SQL 데이터를 Hadoop 으로 로드하는 도구

3. 빅데이터 분석 프로세스 절차

3.2 빅데이터 저장 관리(Big Data Processing)

- 데이터 수집 과정을 통해 확보된 빅데이터로부터 유용한 정보를 추출하려면 빅데이터를 효과적으로 저장 관리하여야 한다.
- 빅데이터 저장이란 검색 수집한 데이터를 분석에 사용하기에 적합한 방식으로 안전하게 영구적인 방법으로 보관하는 것으로서 대용량의 다양한 형식의 데이터를 고성능으로 저장하고 필요한 경우 데이터를 검색하여 수정, 삭제 또는 원하는 내용을 읽어오는 방법을 제공하는 것을 포함한다.
- 빅데이터 저장은 다시 빅데이터 전/후저리와 빅데이터 저장으로 나누어진다.

3. 빅데이터 분석 프로세스 절차

3. 2. 1) 빅데이터 전처리(pre-processing)

- **필터링**: 데이터 활용목적에 맞지 않는 정보는 필터링으로 제거하여 분석시간을 단축하고 저장 공간을 효율적으로 활용하도록 하며 비정형 데이터는 데이터 마이닝을 통해 오류나 중복을 제거하여 저품질 데이터를 개선 처리하는 과정을 말한다. 이때에 자연어처리 및 기계학습과 같은 기술을 적용할 수 있다.

- **유형변환**: 데이터의 유형을 변환하여 분석이 용이한 형태로 변환하는 과정을 말한다.

- **정제**: 수집된 데이터의 불일치성을 교정하기 위한 과정으로 빠진 값(missing value)을 처리하고 데이터 속에 있는 노이즈(noise)를 제거하는 과정을 말한다.

3. 빅데이터 분석 프로세스 절차

3. 2. 2) 빅데이터 후처리(post-processing)

- 데이터 후처리에서의 데이터 변환은 다양한 형식으로 수집 된 데이터를 분석에 용이하도록 일관성 있는 형식으로 변환하는 것을 말하며 **평활화**(smoothing), **집계**(aggregation), **일반화**(generalization), **정규화**(normalization), **속성생성**(attribute/feature construction) **등을 거치게 된다**.

- 데이터 통합은 출처는 다르지만 **상호 연관성이 있는 데이터들을 하나로 결합하는 기술**로 데이터 통합 시 동일한 데이터가 입력 될 수 있으므로 연관관계 분석 등을 통해 중복 데이터를 검출 하거나 **표현 단위(파운드와 kg, inch와 cm, 시간 등)가 다른 것을 표현이 일치하도록 변환**하는 것을 말한다.

- 축소는 분석에 **불필요한 데이터를 축소**하여 고유한 특성은 손상되지 않도록 하고 분석에 대한 효율성을 높이는 과정을 말한다.

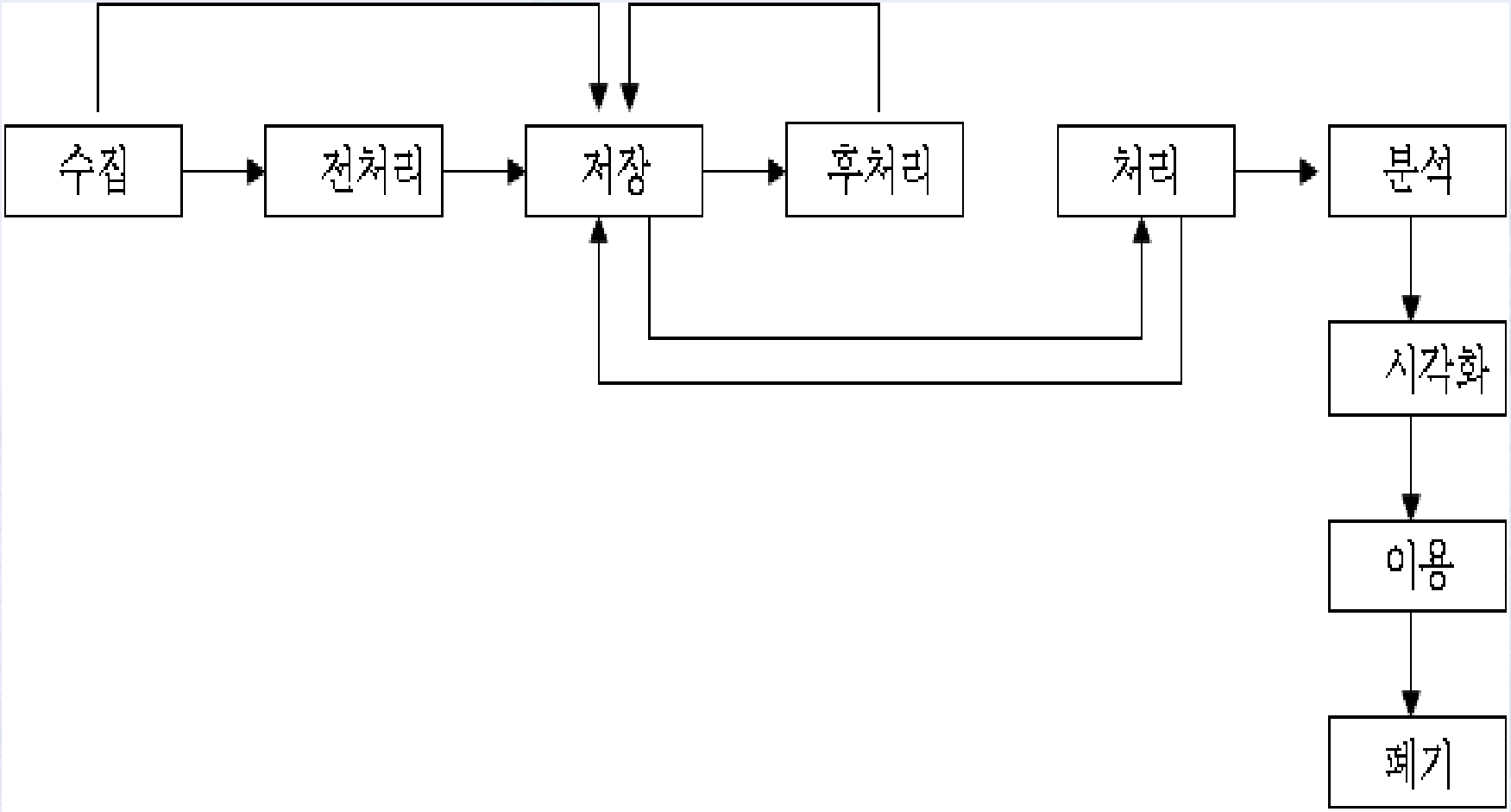
3. 빅데이터 분석 프로세스 절차

데이터 저장 방식의 분류

구분	특징	비고
RDB	-관계형 데이터를 저장하거나, 수정하고 관리할 수 있게 해주는 데이터베이스 -SQL 문장을 통하여 데이터베이스의 생성, 수정 및 검색 등 서비스를 제공	oracle, mssql, mySQL, sybase, MPP DB
NoSQL	-Not-Only SQL의 약자이며, 비관계형 데이터 저장소로, 기존의 전통적인 방식의 관계형 데이터베이스와는 다르게 설계된 데이터베이스 -테이블 스키마(Table Schema)가 고정되지 않고, 테이블 간 조인(Join) 연산을 지원하지 않으며, 수평적 확장(Horizontal Scalability)이 용이 -key-value, Document key-value, column 기반의 NoSQL이 주로 활용 중	MongoDB, Cassandra, HBase, Redis
분산파일시스템	-분산된 서버의 로컬 디스크에 파일을 저장하고 파일의 읽기, 쓰기 등과 같은 연산을 운영체제가 아닌 API를 제공하여 처리하는 파일시스템 -파일 읽기/쓰기 같은 단순연산을 지원하는 대규모 데이터 저장소 -범용 x86서버의 CPU, RAM 등을 사용하므로 장비 증가에 따른 성능 향상 용이 -수 TB ~ 수백PB 이상의 데이터 저장 지원 용이	HDFS(Hadoop File System)

3. 빅데이터 분석 프로세스 절차

빅데이터 분석 프로세스의 데이터 흐름도



3. 빅데이터 분석 프로세스 절차

3.3 빅데이터 처리(Big Data Processing)

- **첫째**, 빅데이터 처리는 기존의 데이터 처리방식과는 다르게 의사결정의 즉시성이 덜 요구된다.
- **둘째**, 대용량의 데이터에 기반을 둔 분석 위주로서 장기적이고 전략적이며 때때로 일회성 거래 처리나 행동 분석을 지원하여야 한다.
- **셋째**, 단순한 프로세싱 모델이 아닌 다양한 데이터 소스, 복잡한 로직 처리, 대용량 데이터 처리 등을 위해 처리의 복잡도가 가장 높고 통상적으로 분산 처리 기술을 필요로 한다.
- **넷째**, 빅데이터는 처리해야 할 데이터양이 방대하고 대용량 처리와 복잡한 처리를 특징으로 하고 있어 실시간 또는 준실시간 처리가 보장돼야 하는 데이터 분석에는 약간 적합하지 않을 수 있다.

3. 빅데이터 분석 프로세스 절차

3.3.1) 빅데이터 일괄처리

- 일괄 처리 기술은 쌓인 빅데이터를 여러 서버로 분산해 각 서버에서 나눠서 처리하고, 이를 다시 모아서 결과를 정리하는 분산, 병렬 기술 방식을 사용한다. 대표적인 기술로는 하둡의 맵리듀스 그리고 마이크로소프트의 드라이애드(Dryad)가 있다.

3. 빅데이터 분석 프로세스 절차

3.3.2) 빅데이터 실시간 처리

빅데이터 처리 관련 기술 분류

대분류	소분류	관련 기술
실시간처리	In-Memory Computing	In-memory 플랫폼, In-memory 메시징, In-memory 데이터관리 (DBMS, Data Grid)
	데이터 스트림 처리	DBMS, Storm, ESPER, S4, Hstreaming CEP (Complex event Processing)
분산처리	Cloud computing	클라우드 컴퓨팅, 분산처리
	Hadoop	HDFS, MapReduce

3. 빅데이터 분석 프로세스 절차

3.4 빅데이터 분석

- 빅데이터로부터 의미 있는 지식을 얻고 이것을 효율적인 의사결정에 활용하려면 빅데이터를 효과적으로 분석할 수 있는 방법과 다양한 인프라가 필요하다.
- 빅데이터 분석은 분석 계획 수립, 분석 시스템 구축, 분석 실행의 3 단계로 구성 된다.

3. 빅데이터 분석 프로세스 절차

빅데이터 분석 소프트웨어 예시

기능	구성요소(예)	주요 내용
빅데이터 수집	Flume. Sqoop. 크롤러. Open API	외부 데이터 추출. 변환. 적재
분산파일 관리	분산파일시스템(HDFS 등)	MapReduce 지원 가능 분산파일 시스템
빅데이터분석	MapReduce	대용량 로그 파일 처리 프레임워크
	Pig	HDFS 대용량 로그 파일을 처리하는 스크립트 언어
	Hive	SQL 기반 대용량 로그 파일의 집계기능을 제공하는 SQL 실행 엔진
	Mahout(머하웃, 마훗)	알고리즘 패키지
	R	오픈소스 통계 패키지

3. 빅데이터 분석 프로세스 절차

3.4.3) 분석 실행

- 빅데이터를 분석하기 위한 기법들은 통계학과 전산학, 특히 기계학습이나 데이터 마이닝 분야에서 이미 사용되던 분석기법들의 알고리즘을 개선하여 빅데이터 분석에 적용시키고 있다. 최근에는 소셜미디어 등 비정형 데이터에 적용이 가능한 텍스트 마이닝, 오피니언 마이닝, 소셜네트워크 분석, 군집분석 등이 주목을 받고 있다. 빅데이터 분석기술의 대표적인 예들로는 빅데이터 통계분석, 데이터 마이닝, 텍스트 마이닝, 예측 분석, 최적화, 평판 분석, 소셜 네트워크 분석, 소셜 빅데이터 분석 등이 있다.

3. 빅데이터 분석 프로세스 절차

3.5 빅데이터 분석 시각화(Visualization)

- 크고 복잡한 빅데이터 속에서 의미 있는 정보와 가치들을 찾아내어 사람들이 쉽게 직관적으로 알 수 있도록 표현하는 기술이 분석 시각화(visualization)이다.

- 시각화의 과정은 Acquire, Parse, Filter, Mine, Represent, Refine, Interact의 7단계로 나누어 설명할 수 있다.

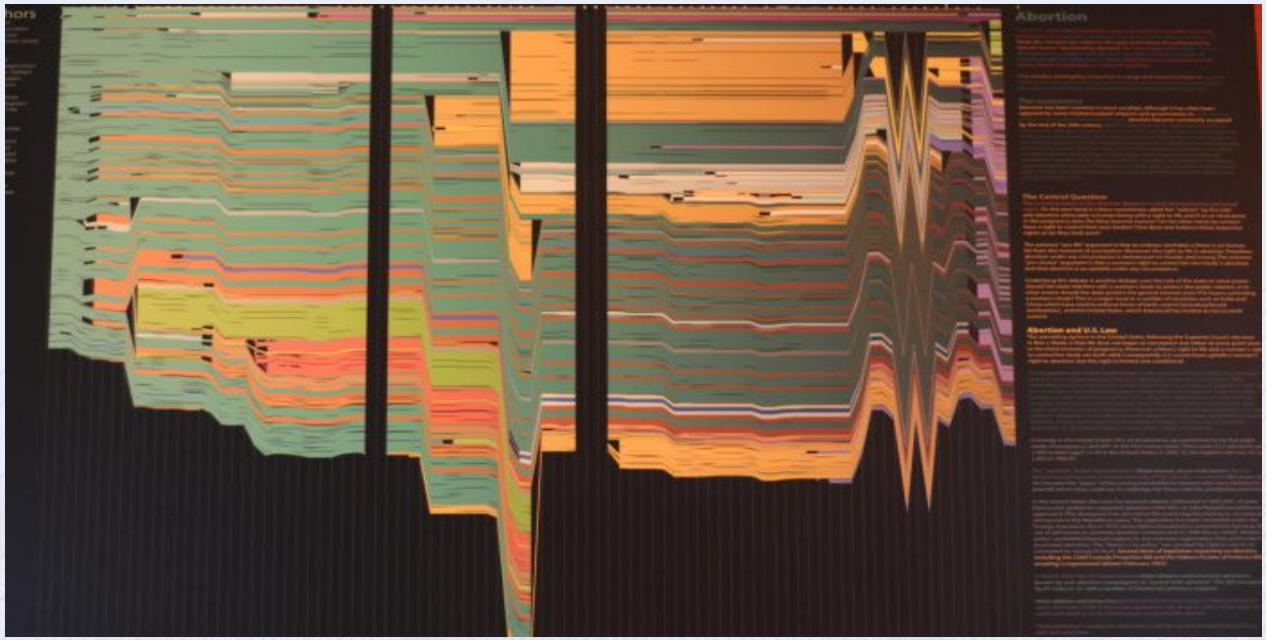
3. 빅데이터 분석 프로세스 절차

3.5 빅데이터 분석 시각화(Visualization)

- Acquire: 디스크의 파일이나 네트워크를 통해서 시각화 하고자 하는 데이터를 획득한다.
- Parse: 데이터의 의미를 해석할 수 있도록 구조에 넣는다.
- Filter: 시각화의 대상이 되는 관심 있는 데이터만 남기고 나머지는 제거한다.
- Mine: 통계학이나 데이터 마이닝 등의 분석 기법을 이용하여 패턴을 파악하거나 수학적인 맥락(mathematical context)을 파악한다.
- Represent: 막대그래프, 리스트(list)나 트리구조(tree) 등의 기본적 시각화 모델을 이용 표현한다.
- Refine: 기본 표상(basic representation)을 더 명확하고 시각적으로 돋보이게 개선시킨다.
- Interact: 사용자가 데이터를 변경하거나 보이는 내용을 조절할 수 있는 방법을 제공한다.

3. 빅데이터 분석 프로세스 절차

시각화의 예: 위키피디아 문서 'abortion(낙태)' 의 히스토리 플로우



- 대중이 문서를 생성하고 개정하는 변화의 모습을 시각적으로 보여줌
- 색깔은 다른 단어를 의미하며 편집자가 늘어나고 글이 개정되면서 수록 띠의 숫자가 늘어남

3. 빅데이터 분석 프로세스 절차

3.6 빅데이터 폐기(Big Data Disposition)

- 빅데이터 폐기 단계에서는 데이터 분석을 위해 이용된 데이터를 삭제하는 단계이며 특히 개인정보와 같은 데이터이거나 또는 정보의 가치가 없는 데이터들은 이용목적을 달성 후 지체 없이 폐기해야 한다(이재식, 2013).

- 그리고 HDFS(Hadoop Distributed File System)같이 데이터를 여러 곳에 복제하여 분산 저장하는 경우에는 모든 데이터의 폐기가 제대로 이루어졌는지를 검증하기 어려운 문제가 있을 수 있다.

4. 결 론

- 빅데이터 분석은 다양한 종류로 이루어진 많은 양의 데이터 속에 숨겨진 패턴이나 알려지지 않은 유용한 정보들을 찾아내기 위하여 데이터를 살펴보는 프로세스로서 데이터 수집, 저장 관리, 처리, 분석 및 지식 시각화, 이용, 폐기의 순환 과정으로 이루어져 있다.
- 빅데이터 분석에서 효과적인 결과를 얻기 위해서는 빅데이터 순환 과정내의 각 단계가 유기적으로 연결되고 통합되어야 하며 각 단계에서는 분석 목적에 맞는 적정한 기술을 사용하여야 한다.