

01

빅데이터 분석 프로세스의 개념



1. 빅데이터 분석 프로세스의 개념



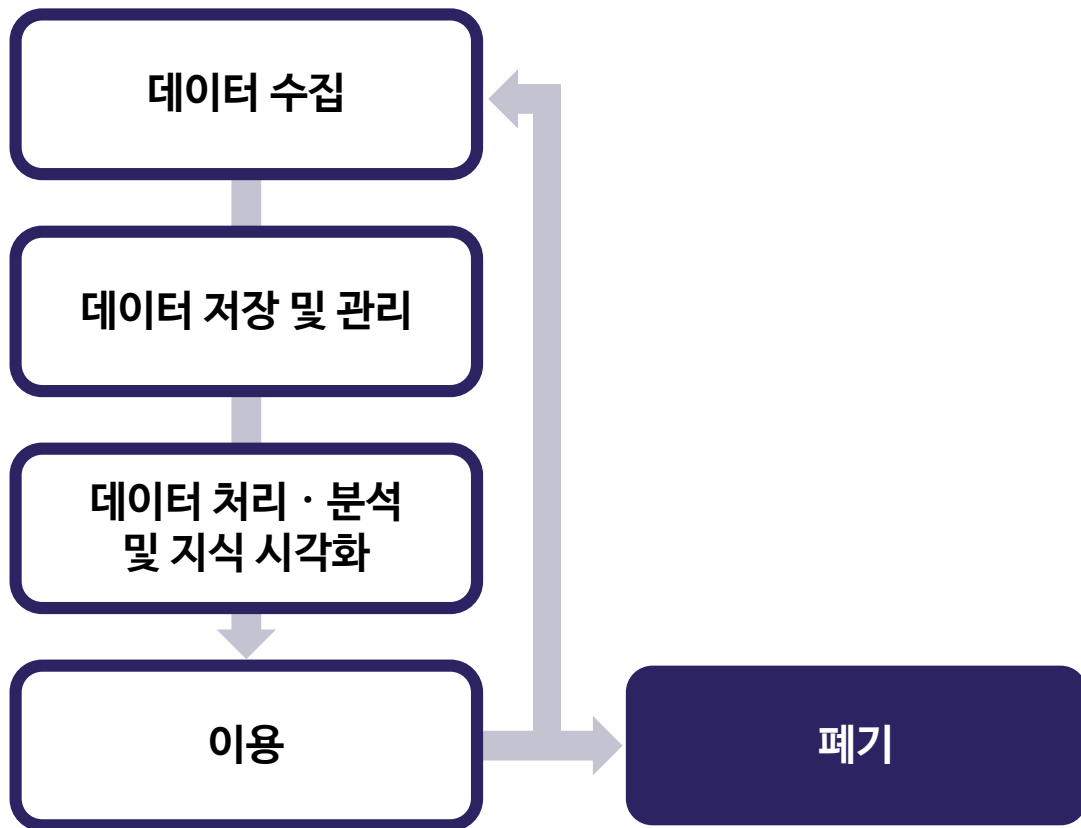
가. 빅데이터 분석의 주요 목적

- 기존의 전통적인 비즈니스 인텔리전스(Business Intelligence) 프로그램이 시도하지 않았던 웹 서버 로그, 인터넷 클릭 정보, 소셜 미디어 활동 보고서, 이동전화 통화 기록, 센서가 감지한 정보 등의 새로운 데이터나 많은 양의 트랜잭션 데이터를 분석하여 기업이 경영과 관련하여 더 좋은 의사결정을 하도록 하는 것

1. 빅데이터 분석 프로세스의 개념



나. 빅데이터 처리의 순환 과정



02

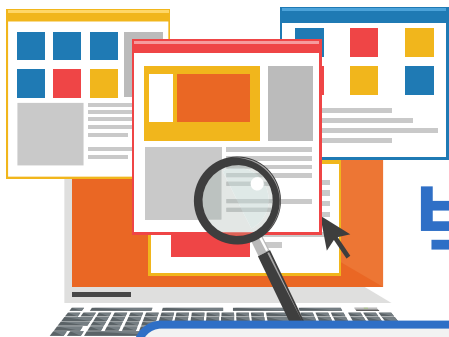
빅데이터 플랫폼



2. 빅데이터 플랫폼



가. 빅데이터 플랫폼의 개념



빅데이터 플랫폼

다양한 데이터 소스에서
수집한 데이터를 분석하여 지식을 추출하고,
이를 기반으로 지능화된 서비스 제공하는 데
필요한 ICT 환경

2. 빅데이터 플랫폼



나. 빅데이터 플랫폼의 능력

빅데이터 처리에 필요한 순환 과정을 수행하기 위해

필요한 **빅데이터 플랫폼 능력**

확장성 있는
대용량 처리 능력

이기종 데이터
수집 및
통합 처리 능력

빠른 데이터 접근
및 처리 능력

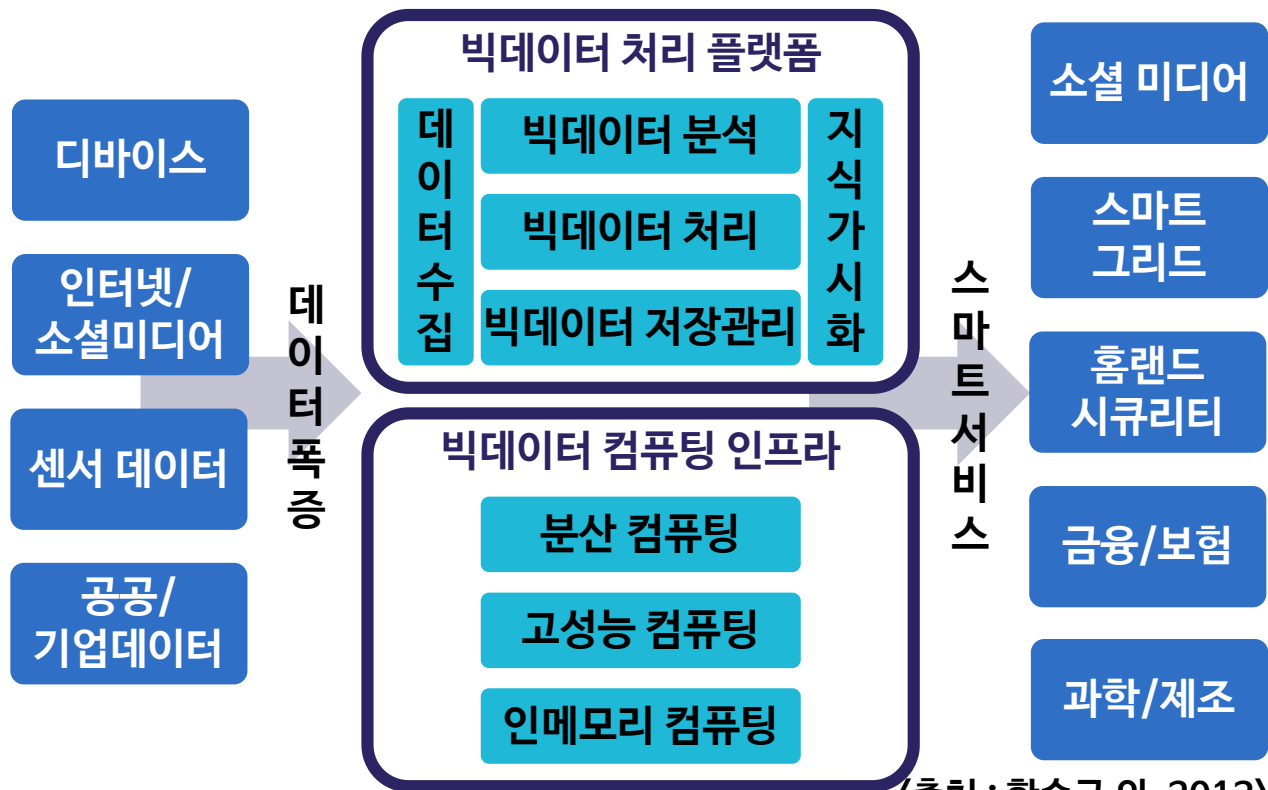
대량의 데이터
저장 관리 능력

대량의
이기종 데이터
분석 능력

2. 빅데이터 플랫폼



다. 빅데이터 플랫폼의 개념도

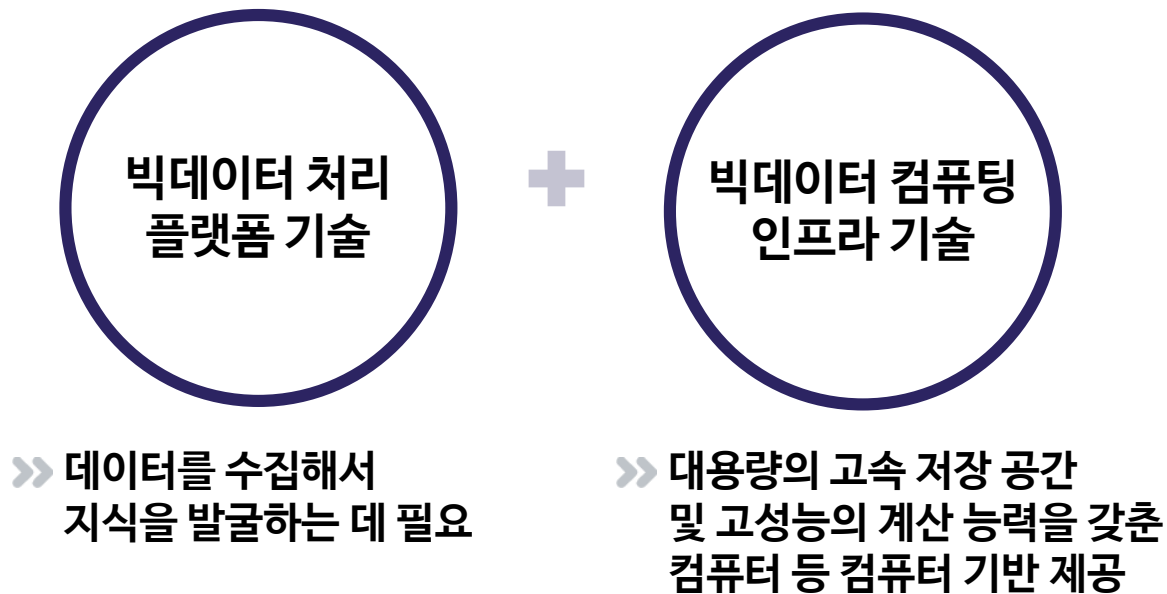


〈출처 : 황승구 외, 2013〉

2. 빅데이터 플랫폼



라. 빅데이터 플랫폼의 구성



2. 빅데이터 플랫폼



라. 빅데이터 플랫폼의 구성

빅데이터 처리 플랫폼

(Big data analytics platform)

다양한 데이터 소스로부터 데이터 수집, 저장 관리,
처리·분석 및 지식 시각화를 통한 지식 이용까지
각 단계를 지원하는 데 필요한 공통 소프트웨어

03

빅데이터 분석 프로세스 절차

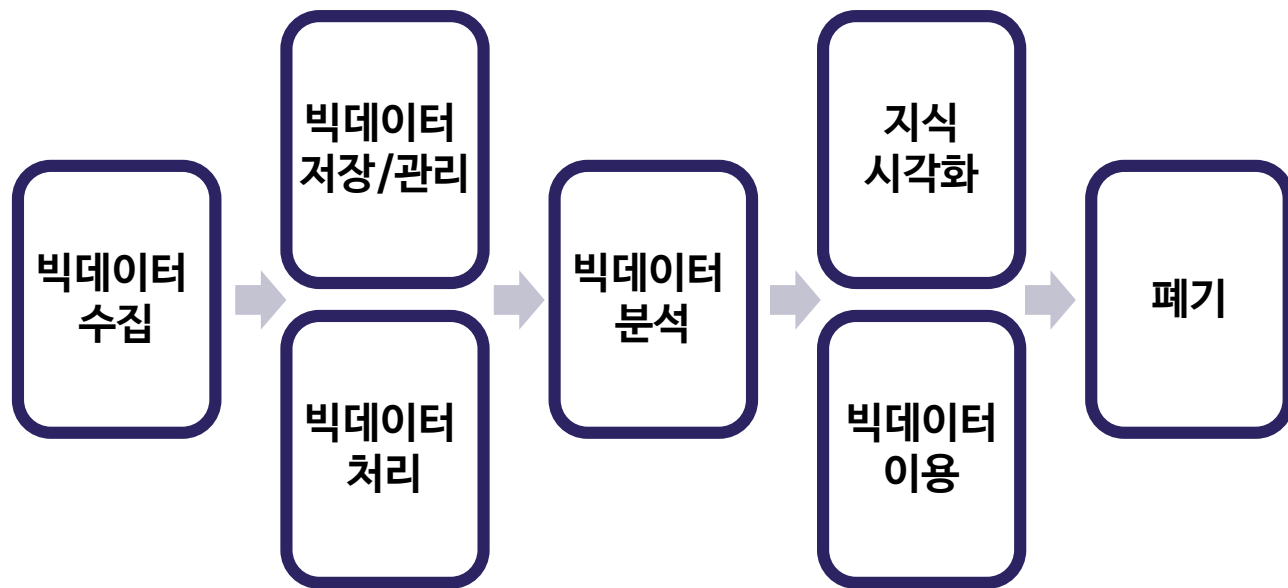
BIG

DATA

3. 빅데이터 분석 프로세스 절차



가. 빅데이터 분석 처리 프로세스



〈출처 : 황승구 외, 2013〉

3. 빅데이터 분석 프로세스 절차



나. 빅데이터 수집 (Big Data Collection)

데이터

- » 기업이나 조직 내부에 있는 정보시스템에 저장된 정형화된 데이터
- » 데이터 수집 시 큰 노력 없이도 가능
- » 수집 후 데이터 가공 시 큰 노력 필요 X

수집하고자 하는 데이터도 개발 단계에서부터 향후에 분석하기에 적합한 정형화된 형식의 로그로 구현하기 때문

빅데이터

- » 조직 내부의 정형화된 데이터와 조직 외부의 무한한 데이터 중 필요로 하는 데이터를 발견 및 수집하고, 정보 분석을 위한 특정 데이터 형식으로 변환하는 과정이 필요

3. 빅데이터 분석 프로세스 절차



나. 빅데이터 수집 (Big Data Collection)

빅데이터 수집

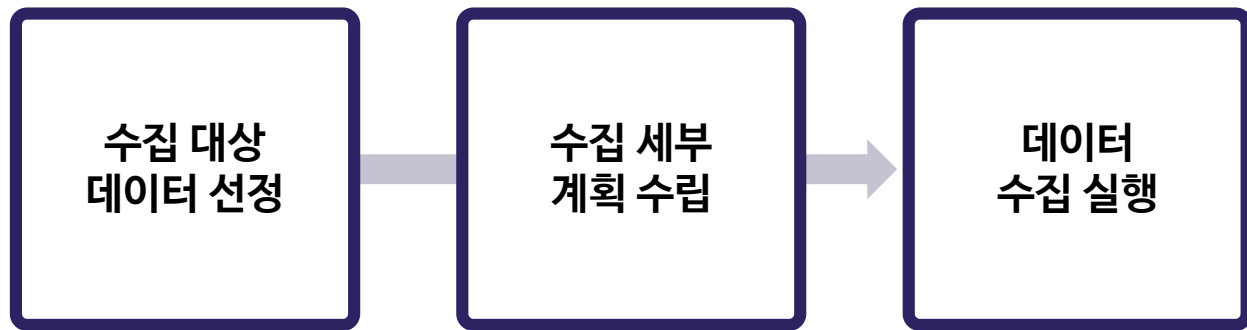
단순히 데이터 확보하는 기술이 아니라,
데이터를 검색하여 수집하고
변환 과정을 통해 정제된 데이터를 확보하는 과정

3. 빅데이터 분석 프로세스 절차



나. 빅데이터 수집 (Big Data Collection)

④ 빅데이터 수집의 세부 절차



〈출처 : 빅데이터 활용 단계별 업무절차 및 기술 활용 매뉴얼 (Version 1.0)〉

3. 빅데이터 분석 프로세스 절차



나. 빅데이터 수집 (Big Data Collection)

01 수집 대상 데이터 선정

- 빅데이터 수집은 빅데이터 분석이나 서비스 제공 시 서비스 품질을 결정하는 중요한 핵심 단계
- 수집 대상 분야에 분석 경험이 많은 전문가 의견을 반영하여 분석 목적에 맞는 데이터를 선정할 것
 - ≫ 대상 데이터의 수집 및 사용 가능 여부 고려
 - ≫ 이용 목적에 맞는 세부 항목 포함 여부 고려
 - ≫ 개인정보 침해 여부나 수집 비용 고려

3. 빅데이터 분석 프로세스 절차



나. 빅데이터 수집 (Big Data Collection)

02 수집 세부 계획 수립

- 몇 가지 사항을 확인하여 적절한 수집 기술을 선정할 것
 - ≫ 데이터 소유자 확인
 - ≫ 대상 데이터가 내부 데이터인지 외부 데이터인지 확인
 - ≫ 수집 대상 데이터의 유형과 데이터 포맷 확인
- 데이터 소스로부터 다양한 데이터를 수집하기 위해 확장성, 안정성, 실시간성, 유연성을 갖춘 기술을 선정할 것

3. 빅데이터 분석 프로세스 절차



나. 빅데이터 수집 (Big Data Collection)

02 수집 세부 계획 수립

☑ 데이터 유형에 따른 수집 기술

데이터 유형	데이터 종류	수집 기술
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

〈출처 : 빅데이터 활용 단계별 업무절차 및 기술 활용 매뉴얼 (Version 1.0)〉

3. 빅데이터 분석 프로세스 절차



나. 빅데이터 수집 (Big Data Collection)

03 데이터 수집 실행

- 데이터를 수집하는 주체의 능동성 여부에 따라
능동적 데이터 수집과 수동적 데이터 수집으로 분류

능동적 데이터 수집

- » 데이터를 가진 주체가 데이터 수집을 원하는 주체에게 능동적으로 데이터 전달
- » 생산설비에서 생산 관련 데이터, 로그데이터 (Log Data), 설문조사 등 제공

수동적 데이터 수집

- » 데이터를 가진 주체가 웹페이지를 통해 데이터를 공개하고, 데이터 수집을 원하는 주체가 웹 로봇, 웹 크롤러 등을 사용하여 웹 페이지에 게시된 정보를 수집

3. 빅데이터 분석 프로세스 절차



나. 빅데이터 수집 (Big Data Collection)

03 데이터 수집 실행

- 데이터 소스의 위치에 따라 내부 데이터 수집과 외부 데이터 수집으로 구분

내부 데이터 수집

- » 주로 자체적으로 보유한 내부 파일 시스템, 데이터베이스 관리 시스템, 센서 등에 접근하여 수집
- » 대표적인 수집 방법은 ETL (Extraction, Transformation, Loading)

외부 데이터 수집

- » 인터넷으로 연결된 외부에서 수집
- » 대표적인 수집 방법은 크롤링 엔진 (Crawling Engine)

3. 빅데이터 분석 프로세스 절차



나. 빅데이터 수집 (Big Data Collection)

04 빅데이터 변환/통합

- 빅데이터 변환은 데이터 수집 과정에서 컴퓨터가 바로 처리할 수 없는 비정형 데이터를 구조적 형태로 전환하여 저장하는 것을 의미
 - ≫ 빅데이터 정제(Cleansing)를 포함
 - ≫ 비정형 데이터를 정제하거나 정형적 데이터에서 빠진 측정값, 다른 형식, 내용 자체가 틀린 데이터를 고쳐주는 과정을 의미
- 데이터 통합은 빅데이터의 효과적인 분석을 위해 레거시 데이터 간 통합을 하고, 비정형 데이터를 수집하는 과정에서 구조적 형태로 전환되어 저장하고, 수집한 비정형 데이터와 레거시 데이터 간의 통합하는 것을 의미


3. 빅데이터 분석 프로세스 절차



나. 빅데이터 수집 (Big Data Collection)

04 빅데이터 변환/통합

④ 빅데이터 수집을 위한 변환 및 통합

ETL (Extraction, Transformation, Load)	메인 프레임, ERP, CRM, Flat file, Excel 파일 등에서 데이터를 추출하여 목표 저장소(DW)의 데이터 형태로 변형한 후 저장
비정형 → 정형	<p>비정형 데이터는 비구조적 데이터 저장소에 저장하거나 어느정도 구조적 형태로 변형 저장</p> <p> Scribe, Flume, chuckwa 등 오픈 소스 솔루션</p>

〈출처 : 김재수, 2012〉

3. 빅데이터 분석 프로세스 절차



나. 빅데이터 수집 (Big Data Collection)

04 빅데이터 변환/통합

④ 빅데이터 수집을 위한 변환 및 통합

레거시
데이터와
비정형
데이터간의
통합

데이터 분석을 위해 수집된 정형의 레거시
데이터와 비정형 데이터간의 통합 필요

Sqoop : RDBMS와 HDFS 간의 데이터를
연결해 주는 기능으로 SQL 데이터를
Hadoop으로 로드하는 도구

3. 빅데이터 분석 프로세스 절차



다. 빅데이터 저장(관리) (Big Data Processing)

빅데이터 저장

- 검색 수집한 데이터를 분석에 사용하기에 적합
- 안전하게 영구적인 방법으로 보관하는 것
- 대용량의 다양한 형식의 데이터를 고성능으로 저장하고, 필요 시 데이터 검색, 수정, 삭제 또는 원하는 내용을 읽어오는 방법 제공까지 포함
- 빅데이터 전/후처리와 빅데이터 저장으로 분류

3. 빅데이터 분석 프로세스 절차



다. 빅데이터 저장(관리) (Big Data Processing)

01 빅데이터 전처리 (Pre-Processing)

- 데이터 수집과 변환 과정에서 **빅데이터 저장소에 적재하기 위해** 수집 데이터를 처리하는 과정

필터링 (Filtering)

- » 필터링을 통해 데이터 활용 목적에 맞지 않는 정보를 제거하여 분석시간을 단축하고 저장 공간을 효율적으로 활용
- » 비정형 데이터는 데이터 마이닝을 통해 오류 및 중복 제거하여 저품질 데이터를 개선 처리
- » 자연어 처리 및 기계학습 같은 기술 적용 가능

3. 빅데이터 분석 프로세스 절차



다. 빅데이터 저장(관리) (Big Data Processing)

01 빅데이터 전처리 (Pre-Processing)

- 데이터 수집과 변환 과정에서 **빅데이터 저장소에 적재하기 위해** 수집 데이터를 처리하는 과정

유형변환
(Transformation)

» 데이터 유형을 변환하여 분석이 용이한 형태로 변환

정제
(Cleansing)

» 수집된 데이터의 불일치성 교정을 위한 과정
» 빠진 값(Missing Value)을 처리하고
데이터 속에 있는 노이즈(Noise) 제거

3. 빅데이터 분석 프로세스 절차



다. 빅데이터 저장(관리) (Big Data Processing)

02 빅데이터 후처리(Post-Processing)

- 저장된 빅데이터를 분석하기 전에 분석에 용이하도록 가공하는 과정

변환
(Transformation)

- » 다양한 형식으로 수집된 데이터를 분석에 용이하도록 일관성 있는 형식으로 변환하는 것
- » 평활화(Smoothing), 집계(Aggregation), 일반화(Generalization), 정규화(Normalization), 속성생성(Attribute/Feature construction)

3. 빅데이터 분석 프로세스 절차



다. 빅데이터 저장(관리) (Big Data Processing)

02 빅데이터 후처리(Post-Processing)

- 저장된 빅데이터를 분석하기 전에 분석에 용이하도록 가공하는 과정

통합 (Integration)

- » 출처는 다르지만 상호 연관성이 있는 데이터들을 하나로 결합하는 기술
- » 데이터 통합 시 동일 데이터가 입력될 수 있으므로 연관관계 분석 등을 통해 중복 데이터를 검출하거나 표현 단위 (파운드와 kg, inch와 cm, 시간 등)가 다른 것을 표현이 일치하도록 변환

3. 빅데이터 분석 프로세스 절차



다. 빅데이터 저장(관리) (Big Data Processing)

02 빅데이터 후처리 (Post-Processing)

- 저장된 빅데이터를 분석하기 전에 분석에 용이하도록 가공하는 과정

축소
(Reduction)

» 분석에 불필요한 데이터를 축소하여
고유한 특성은 손상되지 않도록 하고
분석에 대한 효율성을 높이는 과정

3. 빅데이터 분석 프로세스 절차



다. 빅데이터 저장(관리) (Big Data Processing)

03 빅데이터 저장(Big Data Storage)

- 저장할 데이터의 포맷 등의 유형을 검토하고 데이터 저장 관리에 유리한 저장 방식을 RDB, NoSQL, 분산파일시스템 등으로 선정하여 저장

3. 빅데이터 분석 프로세스 절차



다. 빅데이터 저장(관리) (Big Data Processing)

03 빅데이터 저장(Big Data Storage)

④ 데이터 저장 방식의 분류

구분	특징	비고
RDB	<ul style="list-style-type: none">» 관계형 데이터를 저장하거나, 수정하고 관리할 수 있게 하는 데이터베이스» SQL 문장을 통해 데이터베이스의 생성, 수정 및 검색 등 서비스 제공	oracle, mssql, mySQL, sybase, MPP DB

〈출처 : 빅데이터 활용 단계별 업무절차 및 기술 활용 매뉴얼 (Version 1.0)〉

3. 빅데이터 분석 프로세스 절차



다. 빅데이터 저장(관리) (Big Data Processing)

03 빅데이터 저장(Big Data Storage)

☑ 데이터 저장 방식의 분류

구분	특징	비고
NoSQL (Not-Only SQL)	<ul style="list-style-type: none">» 기존의 전통적인 방식의 관계형 데이터베이스와는 다르게 설계된 데이터베이스» 테이블 스키마가 고정되지 않고, 테이블간 조인(Join) 연산 지원 X» 수평적 확장 용이» key-value, Document key-value, column 기반이 주로 활용 중	Mongo DB, Cassandra, H Base, Redis

〈출처 : 빅데이터 활용 단계별 업무절차 및 기술 활용 매뉴얼 (Version 1.0)〉

3. 빅데이터 분석 프로세스 절차



다. 빅데이터 저장(관리) (Big Data Processing)

03 빅데이터 저장(Big Data Storage)

☑ 데이터 저장 방식의 분류

구분	특징	비고
분산 파일 시스템	<ul style="list-style-type: none">» 분산된 서버의 로컬 디스크에 파일을 저장하고 파일의 읽기, 쓰기 등의 연산을 운영체제가 아닌 API를 제공하여 처리하는 파일 시스템» 파일 읽기, 쓰기 같은 단순연산을 지원하는 대규모 데이터 저장소	HDFS

〈출처 : 빅데이터 활용 단계별 업무절차 및 기술 활용 매뉴얼 (Version 1.0)〉

3. 빅데이터 분석 프로세스 절차



다. 빅데이터 저장(관리) (Big Data Processing)

03 빅데이터 저장(Big Data Storage)

④ 데이터 저장 방식의 분류

구분	특징	비고
분산 파일 시스템	<ul style="list-style-type: none">» 범용 x86서버의 CPU, RAM 등을 사용하므로 장비 증가에 따른 성능 향상 용이» 수 TB~수백 PB 이상의 데이터 저장 지원 용이	HDFS

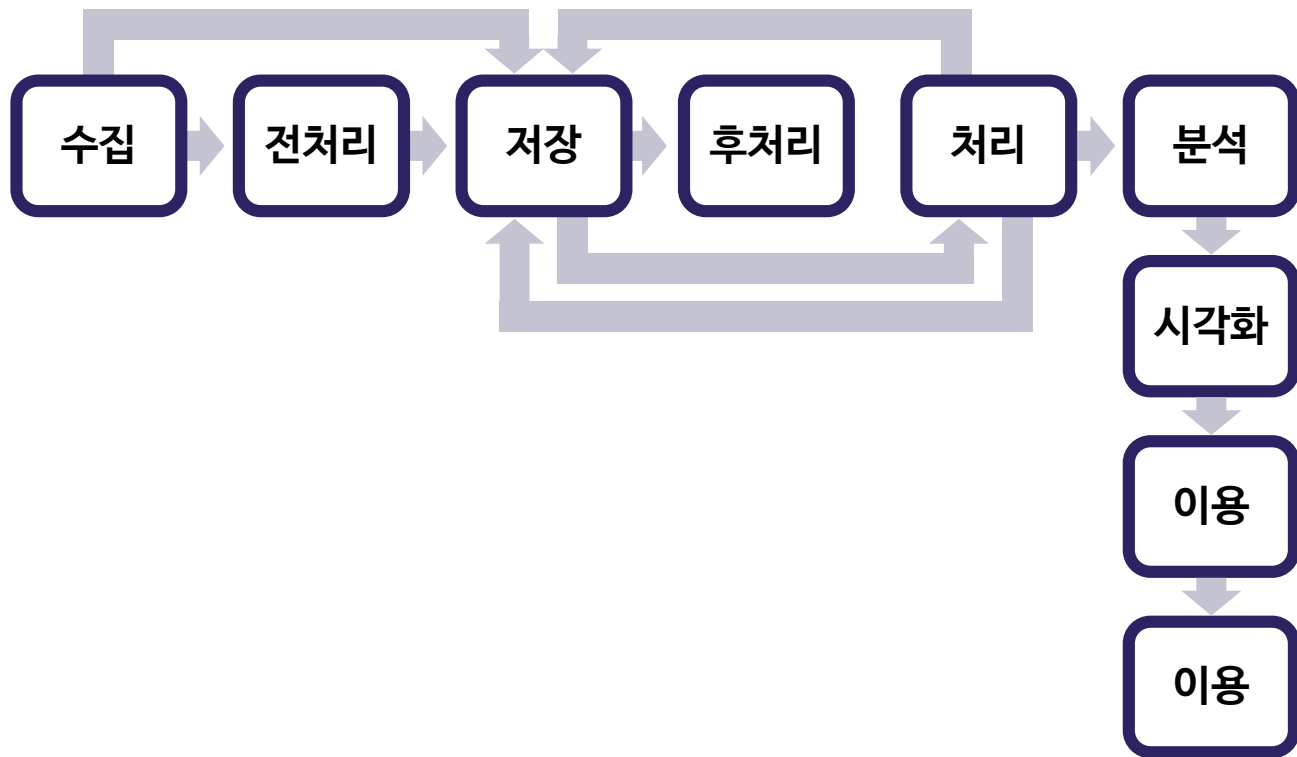
〈출처 : 빅데이터 활용 단계별 업무절차 및 기술 활용 매뉴얼 (Version 1.0)〉

3. 빅데이터 분석 프로세스 절차



다. 빅데이터 저장(관리) (Big Data Processing)

④ 빅데이터 분석 프로세서의 데이터 흐름도



3. 빅데이터 분석 프로세스 절차



라. 빅데이터 처리 (Big Data Processing)

빅데이터 처리

- 빅데이터에서 유용한 정보와 의미 있는 지식을 찾아내기 위한 데이터 가공이나 데이터 분석 과정을 지원하는 과정
- 지속적으로 발생하는 스트림 데이터나 기존 저장소에 저장된 대규모 저장 데이터의 적시 처리를 지원

3. 빅데이터 분석 프로세스 절차



라. 빅데이터 처리 (Big Data Processing)

1

기존 데이터 (CEP, OLTP, ODS, EDW)와는 다르게
의사결정의 즉시성이 덜 요구

3. 빅데이터 분석 프로세스 절차



라. 빅데이터 처리 (Big Data Processing)

- 1 기존 데이터 (CEP, OLTP, ODS, EDW)와는 다르게 의사결정의 즉시성이 덜 요구
- 2 대용량 데이터에 기반을 둔 분석 위주로서 장기적이고 전략적이며 때때로 일회성 거래 처리나 행동 분석을 지원

3. 빅데이터 분석 프로세스 절차



라. 빅데이터 처리 (Big Data Processing)

- 1 기존 데이터(CEP, OLTP, ODS, EDW)와는 다르게 의사결정의 즉시성이 덜 요구
- 2 대용량 데이터에 기반을 둔 분석 위주로서 장기적이고 전략적이며 때때로 일회성 거래 처리나 행동 분석을 지원
- 3 다양한 데이터 소스, 복잡한 로직 처리, 대용량 데이터 처리 등을 위해 처리 복잡도가 가장 높고 통상적으로 분산 처리 기술이 필요

3. 빅데이터 분석 프로세스 절차



라. 빅데이터 처리 (Big Data Processing)

- 1 기존 데이터 (CEP, OLTP, ODS, EDW)와는 다르게 의사결정의 즉시성이 덜 요구
- 2 대용량 데이터에 기반을 둔 분석 위주로서 장기적이고 전략적이며 때때로 일회성 거래 처리나 행동 분석을 지원
- 3 다양한 데이터 소스, 복잡한 로직 처리, 대용량 데이터 처리 등을 위해 처리 복잡도가 가장 높고 통상적으로 분산 처리 기술이 필요
- 4 시간 또는 준실시간 처리가 보장돼야 하는 데이터 분석에는 약간 부적합

3. 빅데이터 분석 프로세스 절차



라. 빅데이터 처리(Big Data Processing)

- 빅데이터 처리는 분산 파일 시스템과 병렬 분산 처리를 이용

» 빅데이터 일괄 처리와 빅데이터 실시간 처리로 구분

빅데이터 일괄 처리

- » 빅데이터를 여러 서버로 분산해 각 서버에서 나눠 처리하고, 다시 모아 결과를 정리하는 분산 · 병렬 기술 방식을 사용
- » 대표적인 기술은 하둡의 맵리듀스, 마이크로소프트의 드라이애드(Dryad)

빅데이터 실시간 처리

- » 통합된 데이터가 엄청난 속도로 생성되는 비정형 데이터 처리를 동시에 효율적으로 하기 위해 필요

3. 빅데이터 분석 프로세스 절차



라. 빅데이터 처리(Big Data Processing)

- 빅데이터 처리는 분산 파일 시스템과 병렬 분산 처리를 이용

» 빅데이터 일괄 처리와 빅데이터 실시간 처리로 구분

대분류	소분류	관련 기술
실시간 처리	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)
분산 처리	Cloud Computing	클라우드 컴퓨팅, 분산처리
	Hadoop	HDFS, MapReduce

3. 빅데이터 분석 프로세스 절차



마. 빅데이터 분석

빅데이터로부터 의미 있는 지식을 얻고
이를 효율적인 의사결정에 활용하려면

빅데이터의 효과적인 분석 방법과 다양한 인프라가 필요



3. 빅데이터 분석 프로세스 절차



마. 빅데이터 분석

01 분석 계획 수립

- 분석을 통해 해결하고자 하는 목적(문제)을 명확히 정의
- 분석 절차와 분석 기업에 대해 세부 시나리오를 작성
- 분석 환경에 대해 분석
 - ≫ 인프라(시스템과 운영 환경)를 자체적으로 기관 내 구축할 것인가?
 - ≫ 외부의 분석 서비스에 위탁을 주고 활용할 것인가?
 - ≫ 자체 인프라와 외부 분석 서비스를 연계하여 활용할 것인가?

3. 빅데이터 분석 프로세스 절차



마. 빅데이터 분석

02 분석 시스템 구축

- 빅데이터 용량이나 분석 작업이 요구하는 부하를 감안하여 수집 데이터 저장 서버, 데이터 처리 서버(하둡 기반 분석, 정형 데이터 DW 등)를 포함하는 **분석 시스템의 하드웨어 인프라를 구축**
- 분석에 필요한 수집, 관리, 분석, 이용자 환경 분석 등 관련 **소프트웨어를 구축**

3. 빅데이터 분석 프로세스 절차



마. 빅데이터 분석

02 분석 시스템 구축

✓ 빅데이터 분석 소프트웨어 예시

기능	구성요소(예)	주요 내용
빅데이터 수집	Flume, Sqoop, 크롤러, Open API	HDFS
분산 파일 관리	분산파일시스템 (HDFS 등)	MapReduce 지원 가능 분산 파일 시스템

3. 빅데이터 분석 프로세스 절차



마. 빅데이터 분석

02 분석 시스템 구축

☑ 빅데이터 분석 소프트웨어 예시

기능	구성요소(예)	주요 내용
빅데이터 분석	MapReduce	대용량 로그 파일 처리 프레임워크
	Pig	HDFS 대용량 로그 파일을 처리하는 스크립트 언어
	Hive	SQL 기반 대용량 로그 파일의 집계기능 제공하는 SQL 실행 엔진
	Mahout	알고리즘 패키지
	R	오픈소스 통계 패키지

3. 빅데이터 분석 프로세스 절차



마. 빅데이터 분석

03 분석 실행

- 빅데이터를 분석하기 위한 기법들은 통계학과 전산학, 특히 기계학습, 데이터 마이닝 분야에서 이미 사용되던 분석기법들의 알고리즘을 개선하여 빅데이터 분석에 적용
- 최근 소셜미디어 등 비정형 데이터에 적용 가능한 텍스트 마이닝, 오피니언 마이닝, 소셜 네트워크 분석, 군집분석 등이 주목
- 대표적인 빅데이터 분석 기술은 빅데이터 통계분석, 데이터 마이닝, 텍스트 마이닝, 예측 분석, 최적화, 평판 분석, 소셜 네트워크 분석, 소셜 빅데이터 분석 등

3. 빅데이터 분석 프로세스 절차



바. 빅데이터 분석 시각화(Visualization)

- 분석 시각화란 크고 복잡한 빅데이터 속에서 의미 있는 정보와 가치들을 찾아내어 사람들이 쉽게 직관적으로 알 수 있도록 표현하는 기술
- 분석한 결과를 활용하여 다양한 시각화 도구로 어떻게 표현하느냐에 따라 직관이 달라지기 때문에 분석 시각화가 중요

3. 빅데이터 분석 프로세스 절차



사. 빅데이터 폐기 (Big Data Disposition)

- 빅데이터 폐기는 데이터 분석을 위해 이용된 데이터를 삭제하는 단계
 - » 특히 개인정보와 같거나 정보 가치가 없는 데이터들은 이용목적을 달성 후 지체 없이 폐기할 것(이재식, 2013)

하드디스크 등
물리적 파기하는 경우

→ 데이터가 저장된 물리적 · 논리적
공간 전체를 폐기하는 방법이어서
일부 데이터만 선택 삭제 어려움

소프트웨어 파기하는 경우

→ 데이터 저장 장소에 다른 데이터를
덮어쓰기 (Overwriting) 작업

HDFS와 같이 데이터 복제로
분산 저장한 것을 폐기하는 경우

→ 모든 데이터의 폐기가
제대로 이루어졌는지 검증 어려움