제3장 빅데이터 수집과 저장

Contents

1. 빅데이터 자료수집

■ 데이터 구축 필요성

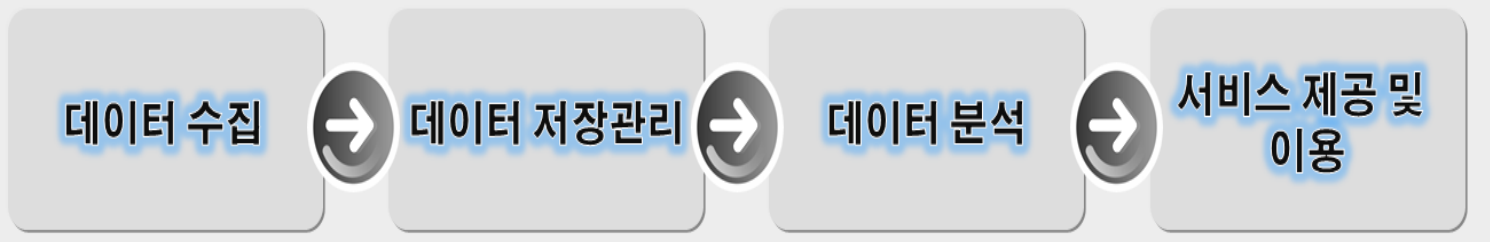
- 스마트폰, SNS, 사물인터넷(M2M) 확산 등에 따른 데이터 폭증
- ICT 인프라 시장 성숙 이후 신규 비즈니스 영역으로 관심 급증
- Davos 포럼에서는 빅데이터를 2012년도의 가장 중요한 기술로 지목
- 주요국 및 글로벌 기업은 빅데이터 산업육성 및 활용에 주력
- 글로벌 ICT기업을 중심으로 빅데이터 핵심기술 및 신규 비즈니스 모델
- 개발이 활발하고, 일반기업의 창의적 활용사례도 속속 등장
- 우리나라는 데이터 생산량이 많은 산업(통신, 조업 등)이 발달하여 잠재력이 크지만, 빅데이터 관련 서비스는 도입 단계임.

1. 빅데이터 자료수집

■ 활용방법 및 범위

1) 데이터 수집과 분석

- 첫째, 데이터 수집은 빅데이터를 수집하는데 필요한 분석절차 및 기술 도입시 고려사항 제시
- 둘째, 데이터 관리는 수집된 데이터를 분석할 수 있도록 가공 및 처리하고 관리하는데 필요한 업무 절차 및 적용기술 제시
- 셋째, 데이터 제공 및 이용은 빅데이터 분석 플랫폼을 이용하여 서비스를 개발하고 분석결과 등을 이용하는데 필요한 절차 및 기술 활용 방안을 제시
- 넷째, 데이터 수집에 있어서 주체에 따른 활용 범위(데이터 제공자, 플랫폼 운영자, 서비스 운영자 등에 따라 활용범위가 각각 다름)



1. 빅데이터 자료수집

■ 활용방법 및 범위

- 데이터 수집에 있어 주체에 따른 활용 범위

주체		활용 범위
데이터 제공자	공공기관 연구기관 민간기업	• 개방 데이터에 대한 범위, 수준, 시기, 방법 등 계획 수립시 참고 • 개인정보 보호, 데이터 보안, 데이터 품질관리 등 정책 수립 및 관리 시 참고 • 데이터제공에 필요한 업무절차, 기술 고려사항 확인
플랫폼 운영자	자체운영 외부제공	• 빅데이터 플랫폼 구축을 위한 설계 및 제안요청서(request or proposal: RFP) 작성에 활용 • 빅데이터 플랫폼 운영(수집, 저장관리, 분석 및 활용 등) • 빅데이터 서비스 개발 참고
서비스 이용자	개발자1인 창조기업 개인 등	• 빅데이터 플랫폼에 접근하여 제공 데이터 이용 및 신규 서비스 개발 등에 활용

1. 빅데이터 자료수집

■ 활용방법 및 범위

• 데이터 접근단계 및 범위

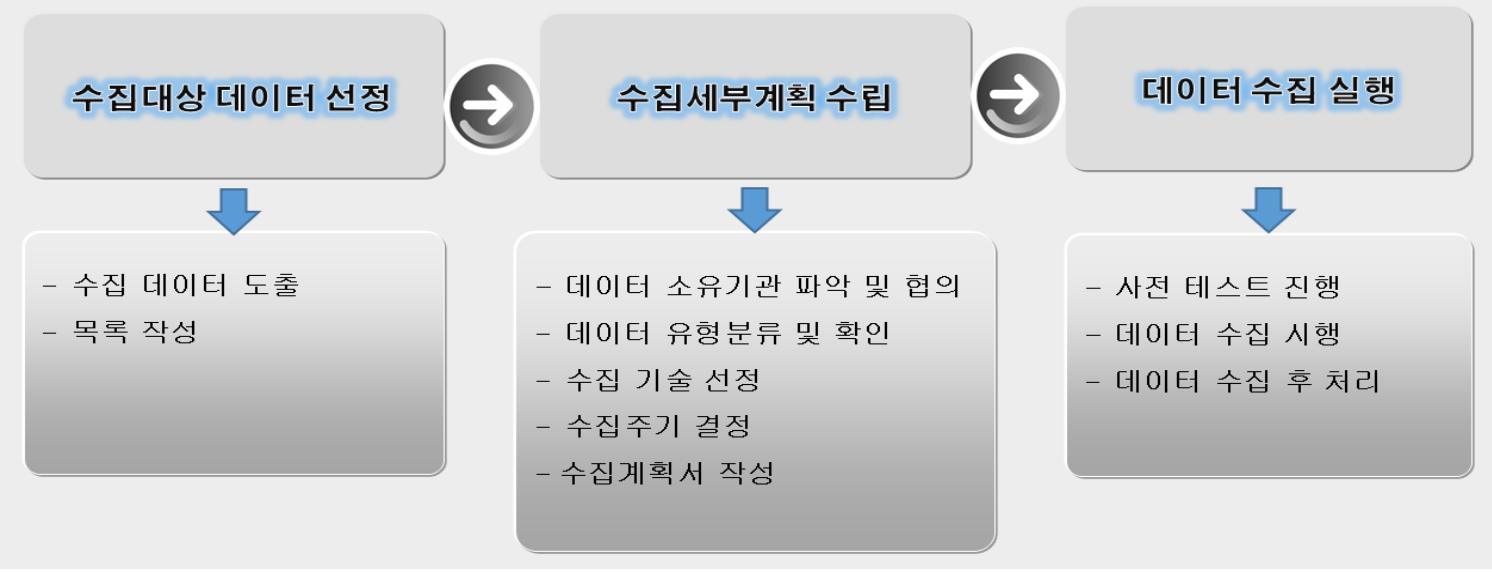
단계	역할	활용 기술	과정
수집	· 내외부 데이터 연동 · 내외부 데이터 통합	Crawing, FTP, Open API, RSS, Log Aggregation, DB Aggregation, Streaming	전처리
적재	· 대용량/실시간 데이터처리 · 분산파일시스템 저장	Distributed File, No-SQL, Memory Cashed, Message Queue	
처리	· 데이터 선택, 변환, 통합, 축소 · 데이터 워크플로 및 자동화	Structured Processing, Unstructured Processing, Workflow, Scheduler	후처리
탐색	· 대화형 데이터 질의 · 탐색적 Ad-Hoc 분석	SQL Like, Distributed Programing, Exploration Visualization	
분석	· 빅데이터 마트 구성 · 통계분석, 고급분석	Data Mining, Machine Learning, Analysis Visualization	
응용	· 보고서 및 시각화 · 분석정보 제공	Data Export/Import, Reporting, Business Visualization	활용

1. 빅데이터 자료수집

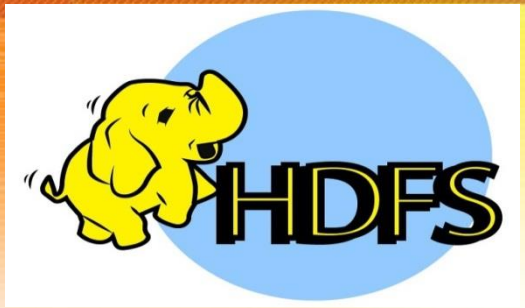
■ 활용방법 및 범위

2) 데이터 수집 절차

- 첫째, 수집대상 데이터 선정은 분석에 필요한 수집 대상 데이터를 선정하되 수집 가능성 여부 등을 파악하고 세부 목록 및 항목을 작성
- 둘째, 세부계획 작성은 수집 데이터 유형을 분류하고 관련 수집기술 및 수집주기, 주요 업무 등을 담은 세부 계획을 작성
- 셋째, 데이터 수집 실행은 수집계획서에 따라 사전 테스트를 진행하여 관련 시스템을 점검한 후 수집활동을 진행



1. 빅데이터 자료수집



Hadoop Distributed File System

■ 활용방법 및 범위

3) 데이터 수집 활용기술

주체	특징	비고
Crawling	• SNS, 뉴스, 웹 정보 등 인터넷상에서 제공되는 웹문서, 정보 수집	• 웹문서 수집
FTP	• TCP/IP 프로토콜을 활용하는 인터넷 서버로부터 각종 파일들을 송수신 • 보안을 강화하기 위해 SFTP 사용 고려 • 서버간 연동시에는 전용 네트워크 구축 고려	• File 수집
Open API	• 서비스, 정보, 데이터 등을 어디서나 쉽게 이용할 수 있도록 개방된 API로 데이터 수집방식 제공 • 다양한 어플리케이션을 개발할 수 있도록 개발자와 사용자에게 공개	• 실시간 데이터 수집
RSS	• RSS(Really Simple Syndication)는 Web기반 최신의 정보를 공유하기 위한 XML 기반 콘텐츠 배급 프로토콜	• 콘텐츠 수집
Strearning	• 인터넷에서 음성, 오디오, 비디오 데이터를 실시간으로 수집할 수 있는 기술	• 실시간 데이터 수집
Log Aggregalor	• 웹서버 로그, 웹 로그, 트랜잭션 로그, 클릭 로그, DB의 로그 등 각종 로그 데이터를 수집하는 오픈 소스 기술 • 종류 : Chukwa, Flume, Scnbe등	• 로그수집
RDB(relation database) Aggregalor	• 관계형 데이터베이스에서 정형 데이터를 수집하여 HDFS(하둡 분산파일 시스템)이나 Hbase와 같은 NoSQL에 저장하는 오픈소스 기술 • 종류 : Sqoop, Direct JDBC/ODBC 등	• RDB 기반 데이터 수집

1. 빅데이터 자료수집

■ 데이터수집 절차

1) 수집 대상 데이터 선정

- 수집 데이터 도출 : 데이터 수집 활동은 데이터 도메인의 분석 노하우가 있는 내 외부 전문가의 의견을 수렴하여 분석 목적에 맞는 데이터를 도출.
- 목록 작성 : 수집 가능성 여부, 보안 문제, 세부 데이터 항목(품질) 및 비용 등을 검토하여 데이터 수집 목록을 작성

가능성	• 해당 데이터가 사용 가능하고 수집 가능한가?
보안	• 수집 시 개인정보 포함여부 및 유출 문제는 없는가?
정확성	• 활용 목적에 따른 세부 항목들이 적절히 포함되었는가?
수집비용	• 데이터 수집에 드는 비용은 얼마인가

1. 빅데이터 자료수집

■ 데이터수집 절차

2) 수집 세부계획 수립

- 데이터 소유기관 파악 및 협의

내부 데이터	• 내부 시스템 간 데이터 연계 가능여부 등을 파악
외부 데이터	• 개방 데이터 종류, 데이터 양, 수집 시스템 연계방식 및 절차, 수집주기 등 관련 기술, 정책을 파악하고 협의
유의사항	• 데이터 수집관련 보안사항, 개인정보보호 관련 문제 등을 점검

1. 빅데이터 자료수집

■ 데이터수집 절차

2) 수집 세부계획 수립

- 데이터 유형 분류 및 확인 : 수집 대상 데이터 유형을 분류하고 데이터 포맷 등 확인, 데이터 유형에는 정형데이터, 반정형 데이터, 비정형 데이터 등 3가지로 구분함

유형	특징	데이터 종류
정형 데이터 (Structured)	· RDBMS의 고정된 필드에 저장 · 데이터 스키마 지원	RDB, 스프레트시트
반정형 데이터 (Semi-structured)	· 데이터 속성인 메타데이터를 가지며, 일반적으로 스토리지에 저장되는 데이터 파일	HTML,XML, JSON, 웹문서, 웹로그, 센서 데이터
	· XML 형태의 데이터로 값과 형식이 다소 일관성이 없음	
비정형 데이터 (Unstructured)	· 언어 분석이 가능한 텍스트 데이터	소셜 데이터, 문서, 이미지, 오디오, 비디오
	· 형태와 구조가 복잡한 이미지, 동영상 같은 멀티미디어 데이터	

1. 빅데이터 자료수집

■ 데이터수집 절차

2) 수집 세부계획 수립

- 수집기술 선정 : 데이터 유형 및 포맷 등에 맞는 수집 기술을 선정, 데이터 유형에 따른 수집 기술로, 반정형 데이터, 비정형 데이터 등에 따라 데이터 종류 정리

데이터 유형	데이터 종류	수집 기술
정형 데이터 (Structured)	• RDB, 스프레드 시트	ETL, FTP, Open API
반정형 데이터 (Semi-structured)	• HTML, XML, JSON, 웹문서, 웹로그, 센서 데이터	Crawling, RSS, Open API, FTP
비정형 데이터 (Unstructured)	• 소셜 데이터, 문서(워드, 한글), 이미지, 오디오, 비디오, IoT	Crawling, RSS, Open API, Streaming FTP

- 수집계획서 작성 : 앞서 소개된 수집대상 '데이터 출처, 수집기술, 수집주기 및 수집 담당자의 주요 업무' 등을 반영하여 계획서를 작성하여 활용

1. 빅데이터 자료수집

■ 데이터수집 절차

3) 데이터 수집 실행

- 사전 테스트 진행 : 수집계획에 따라서 수집주기, 적용기술 등 관련 수집환경에 대한 사전 테스트를 진행

점검 사항	• 네트워크 트래픽 문제, 데이터 누락여부, 정확성(원본 데이터와 샘플 데이터 비교), 보안성(개인정보 포함여부 등) 등을 점검
테스트 수행	• 결과에 따라서 필요시 수집방법 보완 또는 변경
데이터 수집 시행	• 데이터 수집을 진행하되 향후 장애점검 등을 위하여 관련 로그 기록을 확보할 것을 권고
기록	• 수집 데이터의 출처, 수집방식, 장애발생, 로그, 시간 등 수집 당시 상황 등을 시스템적으로 기록

1. 빅데이터 자료수집

■ 데이터수집 절차

3) 데이터 수집 실행

- 데이터 수집 후 처리

- 데이터 수집 후 저장된 데이터에 대한 외부인의 접근 방지 및 유출 시 대처방안 등 관련 업무 지침 마련이 필요

- 빅데이터 서비스 활용을 위한 데이터 수집 시 데이터의 질, 수집 기술, 데이터 보안 및 개인정보 보호 문제 등 다양한 부분을 고려해야 하므로 가급적 관련 전문가의 조언을 받을 것을 권장

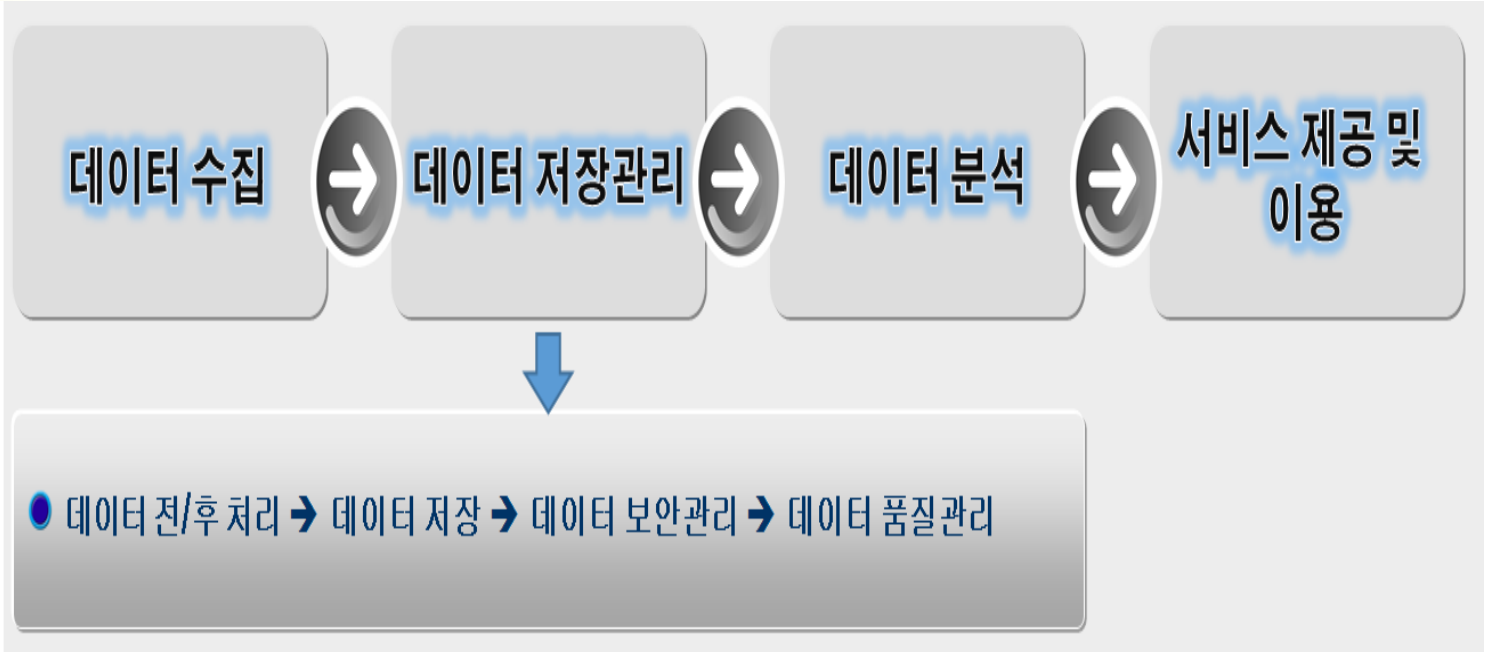
- 데이터 수집 활동은 분석결과의 질을 좌우하는 중요한 과정으로 분석에 필요한 데이터 항목들이 반드시 포함될 수 있도록 사전 점검

- 수집기술은 다양한 데이터 소스로부터 다양한 유형의 데이터를 수집하기 위해 확장성, 안정성, 실시간성 및 유연성 확보가 중요

2. 빅데이터 자료저장

빅데이터 저장관리

- 데이터의 수집과 분석 등의 처리절차는 데이터 수집, 데이터 저장관리, 데이터 분석, 서비스제공 및 이용 등의 절차를 따름
- 일반적인 데이터의 수집과 분석 과정에서 데이터 저장관리는 데이터 처리(전 후) -> 저장 -> 보안 품질관리 등 데이터를 안전하게 활용하기 위해 저장 및 관리 업무를 수행하는 과정



2. 빅데이터 자료저장

■ 데이터 전/후 처리

1) 진행 절차

- 데이터 유형에 따라서 적절한 데이터 처리방식을 선정하여 전/후처리 수행

데이터 처리방식 선정	• 데이터 유형과 분석 목적 등을 고려하여 데이터 저장 전/후처리 기법을 선정
데이터 처리 수행	• 데이터 필터링, 변환, 정제, 통합, 축소 등 선정된 데이터 전/후처리 방식에 따라서 데이터를 처리

2. 빅데이터 자료저장

■ 데이터 전/후 처리

2) 활용 기술

방식	설명
데이터 여과 (Filtering)	• 오류 발견, 보정, 삭제 및 중복성 확인 등의 과정을 통해 데이터 품질을 향상시키는 기술
데이터 변환 (Transformation)	• 데이터 유형 변환 등 데이터 분석이 용이한 형태로 변환하는 기술 • 정규화(Normaliztion), 집합화(Aggreagation), 요약(summarization), 계층 생성 등의 방법 활용 • ETL(Extraction/Transformation/Loading) 도구 제공
데이터 정제 (Cleansing)	• 결측치를 채워 넣고, 이상치를 식별 또는 제거하고, 잡음 섞인 데이터의 불일치성을 교정하는 기술 • 일반적으로 데이터는 불완전하고, 잡음이 섞여있고, 일관성이 없기 때문에 데이터 정제가 필요
데이터 통합 (Integration)	• 데이터 분석이 용이하도록 유사 데이터 및 연계가 필요한 데이터(또는 DB)들을 통합하는 기술
데이터 축소 (Reduction)	• 분석 컴퓨팅 시간을 단축할 수 있도록 데이터 분석에 활용되지 않는 항목 등을 제거하는 기술

2. 빅데이터 자료저장

■ 데이터 전/후 처리

3) 세부 업무처리 절차

- **데이터 처리방식 선정** : 수집된 데이터를 저장하기 위한 전 처리 단계와 저장 된 데이터를 분석하기 전, 후 처리하는 단계로 구분

전 처리	• 수집한 데이터를 저장소에 적재하기 위한 작업으로 데이터 필터링, 유형 변환, 정제 등을 활용
후 처리	• 저장된 데이터를 분석이 용이하도록 가공하는 작업으로 변환, 통합, 축소 등을 활용

2. 빅데이터 자료저장

■ 데이터 전/후 처리

3) 세부 업무처리 절차

- **데이터 유형과 분석 목적** 등을 검토하여 전 후 처리 기술 선택

- 분석에 소요되는 시간과 노력을 절약할 수 있도록 일관성 있는 데이터 형태로 통합

- 분석 효율을 높일 수 있도록 데이터로부터 의미 있는 정보만 추출

- 의미 파악이 어려운 비정형 데이터는 분석이 가능한 형태로 변환

---

2. 빅데이터 자료저장

■ 데이터 전/후 처리

3) 세부 업무처리 절차

- 데이터 처리 수행

- 데이터 중복성, 오류 제거들을 위한 데이터 필터링 기준을 설정

- 실제 사전 테스트를 통하여 오류 발견, 보정, 삭제 및 중복성 검사 등 필터링 과정을 거쳐 필터링 기준을 최적화 하여 활용

- 비정형 데이터는 데이터 마이닝을 통해 오류, 중복, 저품질 데이터를 처리할 수 있도록 자연어처리 및 기계학습과 같은 추가기술 필요

- 분석을 위하여 단위 저장소에 파일형태로 저장 할 경우, 데이터 활용목적에 맞지 않는 정보는 필터링하여 제거해야 분석시간을 단축하고 저장 공간의 효율적 활용이 가능

---

2. 빅데이터 자료저장

■ 데이터 전/후 처리

3) 세부 업무처리 절차

- **데이터 변환** : 다양한 형식으로 수집 된 데이터를 분석에 용이하도록 일관성 있는 형식으로 변환

평활화(Smoothing)	• 데이터로부터 잡음을 제거하기 위해 데이터 추세에 벗어나는 값들을 변환
집계(Aggregation)	• 다양한 차원의 방법으로 데이터를 요약
일반화(Generalization)	• 특정 구간에 분포하는 값으로 스케일을 변화
정규화(Normalization)	• 데이터에 대한 최소-최대 정규화, z-스코어를 말함 • 정규화, 소수 스케일링 등 통계적 기법을 적용
속성 생성 (Attribute/feature construction)	• 데이터 통합을 위해 새로운 속성이나 특징을 만드는 방법으로 주어진 여러 데이터 분포를 대표할 수 있는 새로운 속성/특징을 활용

2. 빅데이터 자료저장

■ 데이터 전/후 처리

3) 세부 업무처리 절차

- **데이터 정제** : 수집된 데이터의 불일치성을 교정하기 위한 방식으로 결측치(Missing Value) 처리와 잡음(Noise) 처리 기술을 활용

방법		설명
결측치	해당 레코드 무시	• 분류에서 스트레스 구분 라벨이 빠진 경우 레코드 무시 • 결측치가 자주 발생하는 환경에서는 적용시 비효율적
	자동으로 채우기	• 결측치에 대한 값을 별도로 정의(예 : “unkown”) • 통계값 적용 : 전체 평균값, 중앙값, 해당 레코드와 같은 클래스에 속한 데이터의 평균값 • 추정치 적용 : 베이지안 확률 추론, 결정 트리
	담당자(전문가) 수작업 입력	• 담당자가 직접 확인하고 적절한 값으로 수정 • 신뢰성은 높을 수 있으나 많은 작업 시간이 소요됨
잡음 (Noise)	구간화(Bining)	• 정렬한 데이터를 여러 개의 구간으로 배분한 후 구간 안에 있는 값들을 대푯값으로 대체 • 구간 단위 별로 잡음 제거 및 데이터 축약 효과 • 사용되는 대푯값 : 평균, Median 등
	회귀값 적용 (Regression)	• 데이터를 가장 잘 표현하는 추세 함수를 찾아서 이 함수의 값을 사용
	군집화(Clustering)	• 비슷한 성격을 가진 클러스터 단위로 묶은 다음 outlier 제거

2. 빅데이터 자료저장

■ 데이터 전/후 처리

3) 세부 업무처리 절차

- **데이터 통합** : 출처가 다른 상호 연관성이 있는 데이터를 하나로 결합하는 기술로 다음 사항들을 고려

- 데이터 통합 시 동일한 데이터가 입력 될 수 있으므로 연관관계 분석 등을 통해 중복 데이터를 검출

- 데이터 통합 전후 수치 및 통계 등 데이터 값들이 일치 할 수 있도록 검증

- 통합 대상 entity가 통합 이후에 동일한지 여부를 확인하기 위한 동일성 검사를 수행

- 표현 단위(파운드와 kg, inch와 cm, 시간 등) 등 서로 다른 방식에 대해 표현을 일치할 수 있도록 변환

---

2. 빅데이터 자료저장

■ 데이터 전/후 처리

3) 세부 업무처리 절차

- **데이터 축소** : 분석에 불필요한 데이터를 축소하여 고유한 특성은 손상되지 않도록 하고 분석에 대한 효율성을 증대시킴

축소 방식		설명
차원 축소	분석에 필요 없거나 중복 항목 제거	• Stepwise forward selection, Stepwise backward elimination 등 활용
데이터 압축	데이터 인코딩이나 변환을 통해 데이터 축소	• Lossless(BMP 포맷) 등 방법 적용
Discrete wavelet transform(DWT)	선형 신호 처리	• 수는 다르지만 길이는 같은 벡터(wavelet coefficients)로 변환 • 여러 개의 벡터 중에서 가장 영향력이 큰 벡터를 선택하여 다른 벡터들을 제거
Principal components analysis(PCA)	데이터를 가장 잘 표현하고 있는 직교 상의 데이터 벡터들을 찾아서 압축	• 속성들을 선택하고 다시 조합시켜 다른 작은 집합으로 생성 • 계산하는 과정이 간단하고 정렬되지 않은 속성들도 처리 가능
수량 축소 (Numerosity Reduction)	데이터를 더 작은 형태로 표현해서 데이터의 크기 줄임	• 데이터 파라미터만 저장(예, Log-linerar 모델) • 기존의 데이터에서 축소된 데이터를 저장(예 : 히스토그램, 클러스터링, 샘플링 등)

2. 빅데이터 자료저장

■ 데이터 전/후 처리

3) 세부 업무처리 절차

- **데이터 전처리**

기능	고려사항
데이터 필터링 (Filtering)	• 데이터 필터링 기준을 정의하고 설정 할 수 있는 기능을 제공해야 함.
	• 데이터 처리 전후에서 생성된 파일의 중복성을 확인 할 수 있도록 파일명, 확장자 등 필터링 기능을 제공해야 함.
	• 유의미한 데이터를 선별하기 위하여 사전 정의된 필터링 기준을 비교 검증 할 수 있는 기능이 제공되어야 함
	• 데이터 필터링 적용시, 비정형 데이터 처리에서 자연어처리 및 기계학습을 수행하기 전에 사용자가 처리 방식을 선택 할 수 있도록 데이터 파일에 대한 정형화된 사전 저장 기준을 제공 하여야 함.
	• 수집된 데이터의 품질 기준의 부합 여부 및 오류 등을 확인하고 관리자에게 알릴 수 있는 기능을 구현해야 함.
	• 필터링 처리 시 사전 정의된 필터링 기준에 의거하여 데이터 처리에서 오류 발생 후 오류에 대한 이력을 저장 할 수 있는 기능을 제공해야 함.

2. 빅데이터 자료저장

■ 데이터 전/후 처리

3) 세부 업무처리 절차

- 데이터 전처리

기능	고려사항
데이터 유형 변환 (Transformation)	• 수집된 데이터의 유형을 분류 할 경우 분류 기준을 적용 할 수 있는 기능을 제공해야 함.
	• 데이터의 유형을 분류하고 이에 대한 데이터 변환에 필요한 알고리즘 함수 또는 변환 구조를 정의 할 수 있는 기능이 제공 되어야 함.
	• 데이터 변환 시 사용자가 지정한 변환 형식에 준하여 변환이 이루어졌는지 확인 할 수 있는 기능이 제공되어야 함.
	• 데이터 변환이 실패 되었을 경우 이력을 저장하고 사용자에게 전달할 수 있는 기능이 제공되어야 함.
	• 데이터 변환이 실패 되었을 경우 이력을 저장하고 사용자에게 전달할 수 있는 기능이 제공되어야 함.
	• 변환된 데이터를 저장하는 기능을 제공해야 함.

2. 빅데이터 자료저장

■ 데이터 전/후 처리

3) 세부 업무처리 절차

- 데이터 전처리

기능	고려사항
데이터 정제 (Cleansing)	• 정제 유형을 사전정의하고 속성 값을 부여하는 기능 및 사용자가 스크립트를 작성 할 수 있는 기능이 제공되어야 함.
	• 데이터 유형별 정제 시 사용자가 설정한 정제 방법을 사전 정의되어 자동으로 지정 할 수 있는 기능이 제공되어야 함.
	• 결측치, 잡음 데이터를 처리하는 경우, 데이터 저장 및 제거 대상에 대하여, 삭제, 처리, 확인 할 수 있는 기능이 제공되어야 함.
	• 데이터의 불일치성을 교정하기 위하여 단위, 표현방식, 코드체계 등의 불일치성을 교정하거나 자동으로 교정이 되도록 하는 자동 스크립팅 기능이 제공되어야 함.

2. 빅데이터 자료저장

■ 데이터 전/후 처리

3) 세부 업무처리 절차

- 데이터 후처리

기능	고려사항
데이터 통합 (Integration)	• 데이터의 일관성을 위해 여러 출처(소스)로부터의 데이터들을 결합할 수 있도록 사전에 확인 할 수 있는 기능을 제공해야 함.
	• 데이터 통합을 위하여 취합된 정보에 대한 상호 관계를 비교하거나 정보 결합 속성 등의 요건을 체크하는 기능이 제공 되어야 함.
	• 데이터 통합 시 통합 전후의 원시 데이터의 백업을 지원하고 이력을 확인 할 수 있는 기능이 제공되어야 함.
	• 데이터 통합을 위해 유일한 키 값을 선정하거나 자동 키(Key) 부여 및 킷값(Key Value) 관리 기능이 제공 되어야 함.

2. 빅데이터 자료저장

■ 데이터 전/후 처리

3) 세부 업무처리 절차

- 데이터 후처리

기능	고려사항
데이터 변환 (Transformation)	• 데이터로부터 잡음을 제거하기 위해 데이터 추세에 벗어나는 데이터(이상치(Outlier)또는 특이값)를 추세에 맞게 변환 또는 자동 추천 할 수 있는 기능을 제공하여야 함.
	• 집계(Aggregation) 시 데이터를 요약하는 기능이 제공 되어야 함.
	• 특정 구간에 분포하는 값을 추출 하거나 이를 사용자가 직관적으로 확인 할 수 있도록 하여 데이터 변환 시 발생 할 수 있는 변환, 패턴, 이벤트를 감시 할 수 있는 기능을 제공해야 함.
	• 데이터 변환 후 사전 저장된 원시 데이터 셋과 변환 후 데이터 간의 변환 로그를 저장 관리 할 수 있는 기능이 제공되어야 함.
데이터 축소 (Reduction)	• 데이터 축소를 위한 적용 기준 또는 적용 스크립트를 부여 할 수 있는 기능이 제공되어야 함.
	• 데이터 크기를 축소하는 경우, 원본 파일의 데이터 축소 범위와 축소가 적용된 속성에 대한 로그를 기록하여 취소 시 재 복구 할 수 있도록 하는 기능이 제공 되어야 함.

2. 빅데이터 자료저장

데이터 저장

- 운영자가 수립 처리된 데이터를 분석에 활용 할 수 있도록 적합한 방식으로 저장 보관하는 작업을 데이터 저장이라 함.
- 데이터 유형에 따라 저장계획을 수립하고 적합한 DB를 구축한 후 데이터 저장 및 관리 함.

저장 계획 수립	• 데이터 유형을 검토하여 저장방식을 선정하고 실행에 필요한 데이터 저장 계획을 수립.
DB 구축 및 테스트 수행	• 선정된 저장방식에 따라 적합한 DB를 구축한 후 사전 테스트를 수행
저장처리 및 모니터링	• 구축된 DB에 데이터를 저장하고 용량한계 등 수시 모니터링 함.

2. 빅데이터 자료저장

■ 데이터 저장

1) 활용 기술

- 데이터 유형에 따라서 데이터 저장 방식은 RDB, NoSQL, 분산파일시스템 등이 있음.

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, NoSQL, 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
분산파일 시스템	• 분산된 서버의 로컬 디스크에 파일을 저장하고 파일의 읽기, 쓰기 등과 같은 연상을 운영체제가 아닌 API를 제공하여 처리하는 파일시스템 • 파일 읽기/쓰기 같은 단순 연산을 지원하는 대규모 데이터 저장소 • 범용 x86서버의 CPU, RAM 등을 사용하므로 장비증가에 따른 성능 향상 용이 • 수 TB~ 수백 PB 이상의 데이터 저장 지원 용이	HDFS

2. 빅데이터 자료저장

■ 데이터 저장

2) 세부 업무처리 절차

- 데이터 유형 검토 : 저장 할 데이터의 포맷 등 유형을 검토하고 데이터 저장관리에 유리한 저장방식을 선정

구분	특징
RDB 저장	• RDB 테이블 데이터는 컬럼과 값을 코드 매핑하거나 데이터형을 변환 처리하여 테이블 형태로 저장 • XML, JSON, HTML 등 형식의 파일은 파싱 처리하여 테이블에 저장 • 문서, 이미지, 비디오, 오디오 등 이진파일은 key값을 추출한 후 테이블에 저장
NoSQL 저장	• 정형데이터(RDB 저장 데이터)는 컬럼과 값을 key와 value로 구분하여 저장 • XML, JSON, HTML 등 형식의 파일은 파싱(parsing)하여 key-value로 저장(NoSQL에서 지원하는 데이터 타입으로 변환 저장) • 문서, 이미지, 비디오, 오디오 파일의 저장은 별도 처리 방안이 필요
분산파일 시스템 저장	• 문서(XML, JSON, HTML, 텍스트 등), 이미지, 비디오, 오디오 등 텍스트 및 이진파일을 분산파일시스템에서 지원하는 파일형식으로 저장

2. 빅데이터 자료저장

■ 데이터 저장

2) 세부 업무처리 절차

- 저장 공간 용량 설계 : 수집 할 데이터 크기 및 최대 저장기간 등을 고려하여 용량 설계

구분	특징	Scale out
개요	• CPU, 메모리, 하드디스크 등 서버 자원을 추가하여 처리 능력을 향상시키는 방식	• 서버의 대수(노드)를 추가하여 처리 능력을 향상 시키는 방식
비용	• 컨트롤러나 네트워크 인프라 비용은 발생하지 않고 디스크만 추가	• 추가된 노드들이 하나의 시스템으로 운영되기 위한 NW장비 필요
용량	• 하나의 스토리지 컨트롤러가 지원 가능한 Device 수가 한정되어 있어 용량 확장시 제약	• 스토리지 용량 확장성이 매우 좋음

2. 빅데이터 자료저장

■ 데이터 저장

2) 세부 업무처리 절차

- DB구축 및 테스트 수행 :

  데이터 저장계획서에 따라서 확장성 등을 고려하여 DB를 구축하고 운영에 필요한 주요기능에 대한 사전테스트를 진행

- 저장처리 및 모니터링 수행

  - 시, 일, 주, 월 주기적으로 데이터 운영관련 오류 및 여유 공간 등을 신시간 모니터링하고 문제 발생 시 신속한 대응체계를 마련

  - 오류가 발생하였을 경우, 내역을 분석하여 수집, 처리, 저장 단계 관련 담당자와 협의하여 해결

3. 빅데이터 자료저장 방법

■ 데이터 베이스 관리 시스템

1) 데이터베이스 관리 시스템(DBMS; DataBase Management System)

- 사용자와 응용프로그램과 데이터베이스간의 인터페이스 역할을 담당하여 데이터 베이스를 응용 프로그램들이 직접 조작하는 것이 아닌, 데이터베이스 조작을 수행하는 별도의 소프트웨어가 있는데 이를 데이터베이스 관리 시스템 이라 함.

기능	내용
정의 기능	• 데이터베이스와 응용 프로그램 간의 상호 작용 수단을 제공 • 물리적 저장 장치에 데이터베이스가 저장될 수 있게 무리적인 구조를 정의
조작 기능	• 데이터베이스와 사용자 간 상호 작용 수단(데이터 요청, 변경 등)을 제공 • 데이터의 처리를 위한 데이터의 삽입, 삭제, 검색, 갱신 등을 지원
제어 기능	• 데이터 간의 모순성이 발생하지 않도록 함 • 데이터베이스의 내용을 항상 정확하게 유지하여 데이터의 무결성이 파괴되지 않도록 함

3. 빅데이터 자료저장 방법

■ 데이터 베이스 관리 시스템 소개

2) DBMS의 장단점

장점	• 데이터 공유, 데이터의 중복을 최소화할 수 있음 • 데이터 중복 감소, 데이터를 공용할 수 있음 • 데이터 일치, 데이터의 일관성을 유지할 수 있음 • 데이터 무결성 유지, 데이터의 무결성을 유지할 수 있음 • 데이터 보안 유지, 데이터의 보안을 보장할 수 있음 • 데이터 표준화 기능, 전체 데이터의 요구사항을 파악하여 조정할 수 있음
단점	• 과다한 비용의 지출이 발생 • 상대적으로 성능이 저하될 수 있음

3. 빅데이터 자료저장 방법

■ 관계형 데이터베이스

1) **관계형 데이터베이스**(Relational Database)

- 메타 데이터를 총괄 관리할 수 있기 때문에 데이터의 성격, 속성 또는 표현 방법 등을 체계화할 수 있음.

- 데이터 표준화를 통한 데이터 품질을 확보할 수 있음.

- DBMS는 인증된 사용자만이 참조할 수 있도록 보안기능을 제공.

- 데이터 무결성을 보장할 수 있음.

- 관계형 데이터베이스의 여러 장점이 알려지면 서 기존의 파일시스템과 계층형, 망형 데이터베이스를 대부분 대체하면서 주력 데이터베이스가 됨.

3. 빅데이터 자료저장 방법

관계형 데이터베이스

2) 관계형 데이터베이스 특징

- 데이터 간의 관계를 표현하기 위해 테이블 집합을 사용.
- 테이블이라는 단순한 개념 및 견고한 수학적 토대를 갖고 있음.
- 프로그래머가 이해하기가 수월하고 실제 데이터베이스시스템을 구현하는데 용이.
- 테이블들의 모임으로 구성되며, 각 테이블은 고유한 이름을 갖음.
- 다른 데이터베이스 모델에 비해 수정이 용이.
- 다른 유형의 데이터베이스 구조를 관계 데이터베이스로 변환하는 일이 비교적 수월함.
- 민감한 데이터에 대한 엑세스 제어를 구현하기 쉬움.
- 엑세스 할 때 상당히 조심해야 하는 데이터는 별도로 관계를 정의하여 두고, 그 관계를 엑세스 할 수 있는 권한을 두거나 또는 액세스 기법을 따로 두는 제어가 가능함.
- 데이터베이스의 간결성(Clarity)과 가시성(Visibility)이 증진됨.
- 포인터를 사용하여 복잡하게 연관되어 있는 데이터 요소들에 대한 탐색보다는 표 형식의 데이터 요소를 탐색하는 편이 훨씬 수월함.

3. 빅데이터 자료저장 방법

관계형 데이터베이스

3) RDB의 종류

- RDBMS(Relational Database Management System) : RDBMS는 관계형 데이터베이스를 생성하고 수정하고 관리할 수 있는 소프트웨어임.

종류	장점	단점
Oracle	• 분산처리 지원기능이 우수함 • SMP 및 MPP의 지원이 가능	• DBMS관리가 복잡 • 가격이 DBMS보다 비쌈
Microsoft SQL Sever	• 윈도우 환경에서 가장 많이 사용됨 • 제품가격이 저렴함	• 충분한 지원 및 발전이 불투명함
My SQL	• 빠르고 안정적이며 사용하기 쉬움 • 고사양을 요구하지 않음	• 실시간 백업이 안됨
Sybase	• 하드웨어 자원만으로도 충분히 활용가능	• 자체 개발등의 지원도구가 부족
Informix	• 사용자들의 만족도 우수 • 낮은 사양에서도 운영 우수 • 지원이 풍부	• PC급 지원기능의 한계가 있음
MS Access	• 확장이 용이	• 윈도우에서만 사용 가능

3. 빅데이터 자료저장 방법

분산 데이터베이스

1) 분산 데이터베이스 개요

- 분산 데이터베이스의 정의는 "여러 곳으로 분산되어 있는 데이터베이스를 하나의 가상 시스템으로 사용할 수 있도록 한 데이터베이스 " 를 뜻함

- 논리적으로 동일한 시스템에 속하지만 컴퓨터 네트워크를 통해 물리적으로 분산되어 있는 데이터들의 모임.

- 물리적 Site 분산: 논리적으로 사용자 통합, 공유, 즉, 분산 데이터베이스는 데이터베이스를 연결하는 빠른 네트워크 환경을 이용하여 데이터베이스를 여러 지역 여러 노드로 위치시켜 상용성, 성능 등을 극대화 시킨 데이터베이스라고 정의할 수 있음.

3. 빅데이터 자료저장 방법

분산 데이터베이스

2) 분산 데이터베이스의 적용 방법 및 장단점

- 분산 데이터베이스 적용방법 : 분산 환경의 데이터베이스를 성능이 우수하게 현장에서 가치 있게 사용하는 방법은 업무의 흐름을 보고 업무구성에 따른 아키텍처 특징에 따라 데이터베이스를 구성하는 것.

- 분산 데이터베이스 장단점

장점	단점
• 지역 자치성, 점증적 시스템 용량 • 신뢰성과 가용성 • 효용성과 융통성 • 빠른 응답 속도와 통신비용 절감 • 데이터의 가용성과 신뢰성 증가 • 시스템 규모의 적절한 조절 • 각 지역 사용자의 요구 수용 증대	• 소프트웨어 개발 비용 • 오류의 잠재성 증대 • 처리 비용의 증대 • 설계, 관리의 복잡성과 비용 • 불규칙한 응답 속도 • 통제의 어려움 • 데이터 무결성에 대한 위협

3. 빅데이터 자료저장 방법

분산 데이터베이스

3) 구글 파일 시스템

- 구글 파일 시스템(GFS)은 급속히 늘어나는 구글의 데이터 처리량을 해결하기 위해서 설계됨.
- GFS는 현재 그리고 예측 가능한 미래의 어플리케이션 부하와 기술 환경에 대한 관측을 통해서 전통적인 방식들을 재검토하고 설계에 대해서 근본적으로 다른 시점으로 설계함.
- 기존에 비해 훨씬 증가한 파일들의 크기를 고려함.
- 대부분의 파일이 새로운 데이터에 의해 덮어지는 것이 아니라 추가확장 되어간다는 점을 고려함.
- 어플리케이션과 파일 시스템 API를 같이 설계한다는 것은 유연성 측면에서 전체 시스템의 효율을 높여줌.
- 현재 다양한 GFS 클러스터들이 서로 다른 목적을 위해 구성되어 있으며, 가장 큰 클러스터는 1000개 이상의 저장 노드와 300테라 이상의 디스크 저장영역을 가지고 서로 다른 기기에서 수백의 클라이언트에게 끊임없이 이용되고 있음.

3. 빅데이터 자료저장 방법

분산 데이터베이스

4) 하둡(Hadoop)

- 빅데이터 시대를 출현시켰던 요인 중 하나가 바로 하둡임.
- 빅데이터는 대개 전통적인 데이터베이스(DB)나 시스템 환경에서 처리하기 힘든 대용량 데이터를 저장, 분석, 처리해 가치 있는 정보로 만들어내는 일련의 과정을 뜻함.
- 대용량 데이터를 처리하는데 공통점은 바로 하둡을 이용해 처리한다는 점.
- 하둡은 국내외를 막론하고 빅데이터를 다루는 개발자들의 관심을 받고 있음.
- 전문가들은 하둡 생태계를 통해 빅데이터를 보다 원활하고 효율적으로 분석할 수 있다고 봄.

3. 빅데이터 자료저장 방법

분석용 데이터베이스 구축

1) Data, DB, DBMS의 개념적 구분

구분	내용
데이터 (Data)	• 인간이 사물을 인식하여 의사소통이 가능한 형태로 표현해 놓을 것을 말함 • 자료는 문자, 숫자, 이미지, 사운드 등 다양한 형태로 표현됨 • 이러한 자료를 전자적 형태로 표현해야 컴퓨터에서 처리가 가능 • 저자적 형태로 표현한 자료는 디지털자료라고 해야 정확한 표현이지만, 일반적으로 현대의 기업환경에서 자료라고 하면 디지털 자료를 의미
데이터베이스 (Database)	• 한 조직 내에서 관련된 자료들을 정보 생산을 목적으로 논리적 관계에 따라 분류하고 정리해서 전자적 매체에 저장해 높은 것을 말함 • 목적지향적이고 공유를 전제로 하고, 데이터 간에 상호 밀접한 관계를 가지고 있어야 함
데이터베이스 관리시스템 (Data Base Management System)	• 사용자의 권한을 체크하고 데이터 간의 상호 연관성을 점검해 가며, 자료의 저장, 변경, 삭제, 검색 등의 작업을 수행하고, 그 결과를 응용프로그램에 전달함.

3. 빅데이터 자료저장 방법

■ 분석용 데이터베이스 구축

2) 데이터웨어하우스의 개념

- 계속 증가되는 기업의 데이터를 정보시스템부서의 전문가뿐만 아니라 해당 업무의 현업 전문가들도 정확히 이해하고 이들 데이터를 이용하여 신속하고도 정확한 의사결정을 하는것이 중요함.
- 의사결정을 위해서 사용되는 데이터를 운영데이터와 구분하여 정보데이터라고 하며, 데이터웨어하우스는 정보데이터를 효과적으로 관리하는 새로운 기술임.

구분	운영데이터	정보데이터
데이터 내용	• 현재 값	• 요약된 값, 과거 값, 계산된 값
데이터 조직	• 응용 시스템별로 조직	• 주제별로 조직
데이터 안정성	• 트랜잭션 발생 때 마다 변동	• 일괄수정 될 때까지 불변
데이터 구조	• 트랜잭션 처리를 위해서 최적화	• 복잡한 질의를 위해서 최적화
접속 빈도	• 매우 빈번	• 가끔
접속 유형	• 검색/수정/삭제	• 검색/총계
사용	• 예측업무 • 반복적 업무	• 비구조적 업무 • 특별한 업무
반응속도	• 3초 내의 짧은 시간에 처리완료	• 수초에서 수분 내에 처리완료